Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **502,** 3456–3475 (2021) doi:10.1093/mnras/staa3774
Advance Access publication 2020 December 15

COSMIC BIRTH: efficient Bayesian inference of the evolving cosmic web from galaxy surveys

Francisco-Shu Kitaura[1,2★] Metin Ata[3] Sergio A. Rodríguez-Torres,[1,2] Mónica Hernández-Sánchez,[1,2] A. Balaguera-Antolínez[1,2] and Gustavo Yepes[4,5]

[1] *Instituto de Astrofísica de Canarias (IAC), Calle Vía Lactea s/n, E-38200 La Laguna, Tenerife, Spain*
[2] *Departamento de Astrofísica, Universidad de La Laguna (ULL), E-38206 La Laguna, Tenerife, Spain*
[3] *Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwa, Chiba 277-8583, Japan*
[4] *Departamento de Física Teórica, Módulo 8, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*
[5] *CIAFF, Facultad de Ciencias, Universidad Autónoma de Madrid, E-28049 Madrid, Spain*

Accepted 2020 December 2. Received 2020 November 20; in original form 2019 November 1

ABSTRACT

We present COSMIC BIRTH (COSMological Initial Conditions from Bayesian Inference Reconstructions with THeoretical models): an algorithm to reconstruct the primordial and evolved cosmic density fields from galaxy surveys on the light-cone. The displacement and peculiar velocity fields are obtained from forward modelling at different redshift snapshots given some initial cosmic density field within a Gibbs-sampling scheme. This allows us to map galaxies, observed in a light-cone, to a single high redshift and hereby provide tracers and the corresponding survey completeness in Lagrangian space including tetrahedral tessellation mapping. These Lagrangian tracers in turn permit us to efficiently obtain the primordial density field, making the COSMIC BIRTH code general to any structure formation model. Our tests are restricted for the time being to augmented Lagrangian perturbation theory. We show how to robustly compute the non-linear Lagrangian bias from clustering measurements in a numerical way, enabling us to get unbiased dark matter field reconstructions at initial cosmic times. We also show that we can accurately recover the information of the dark matter field from the galaxy distribution based on a detailed simulation. Novel key ingredients to this approach are a higher order Hamiltonian-sampling technique and a non-diagonal Hamiltonian mass matrix. This technique could be used to study the Eulerian galaxy bias from galaxy surveys and could become an ideal baryon acoustic reconstruction technique. In summary, this method represents a general reconstruction technique, including in a self-consistent way a survey mask, non-linear and non-local bias, and redshift-space distortions, with an efficiency about 10 times superior to previous comparable methods.

Key words: methods: analytical – methods: statistical – galaxies: distances and redshifts – large-scale structure of Universe – cosmology: observations.

1 INTRODUCTION

The observed accelerated expansion of the Universe (Riess et al. 1998; Perlmutter et al. 1999) poses some of the most intriguing questions in modern cosmology: what is the origin of such a dynamical state? (see e.g. Guzzo et al. 2008); is the so-called dark energy component responsible for it?; and what is its nature? In the recent years, a number of wide-field galaxy surveys have been designed in order to answer these fundamental questions, such as Extended Baryon Oscillation Spectroscopic Survey (eBOSS; Dawson et al. 2016), *Euclid* (Amendola et al. 2018), Dark Energy Spectroscopic Instrument (DESI; Levi et al. 2013), 4-metre Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019), *Wide-Field Infrared Survey Telescope* (*WFIRST*; Akeson et al. 2019), and Large Synoptic Survey Telescope (LSST; LSST Science Collaboration 2009). Additionally, pencil-beam surveys with smaller footprints but deeper and more abundant target sampling, e.g. VIMOS Public Extragalactic Redshift Survey (VIPERS; Guzzo et al. 2014) and Prime Focus Spectrograph (PFS; Takada et al. 2014), and also dense and deep photometric surveys, such as Dark Energy Survey (DES; The Dark Energy Survey Collaboration 2005) and Javalambre-Physics of the Accelerated Universe Astrophysical Survey (J-PAS; Benitez et al. 2014), contribute to understand the cosmic evolution of large-scale structures. The acquisition of observational data, as generated by these surveys, is potentially reaching the precision requirements to be able to tackle not only the above-mentioned questions, but indeed more profound ones, such as the validity of general relativity on the largest cosmological scales (see e.g. Ishak 2019, and references therein). To that end, the tools envisaged to perform data analysis need to keep track with the observational campaigns in order to be able to exploit the data to its maximal information content.

Recent measurements of the statistical properties of the spatial distribution of galaxies (see e.g. Alam et al. 2017) or quasars (see e.g. Ata et al. 2018) are currently paving the road for careful analyses aiming at shedding light into the different physical processes involved in the formation of large-scale structures. The identification of some

★ E-mail: fkitaura@iac.es

of such properties dates back to pioneering papers (see Shandarin & Zeldovich 1989; Bond, Kofman & Pogosyan 1996; Platen et al. 2011), with the realization that galaxies follow an intricate pattern, the *cosmic web*. A large variety of tools have been envisaged to study it (see e.g. Libeskind et al. 2018, and references therein). Another particular feature in the spatial distribution of galaxies is the baryon acoustic oscillations (BAO). Its detection (see e.g. Eisenstein et al. 2005; Percival et al. 2007) represents the establishment of cosmological standard ruler, and a sensitive probe for the equation of state of the dark energy (e.g. Blake & Glazebrook 2003; Seo & Eisenstein 2003; Aubourg et al. 2015; Zhao et al. 2017).

Extracting the content of cosmological information encoded in the BAO is not a trivial task. At early cosmological times, such information was entirely contained in the two-point correlation function (or its Fourier counterpart, the power spectrum), which fully characterizes linear Gaussian overdensity fields. However, as cosmic density fields evolved, non-linear evolution not only induced shifts in the position of the acoustic peak (e.g. Crocce & Scoccimarro 2008), but also dragged part of the information of the BAO signature from the two-point to higher order statistics (see e.g. Schmittfull et al. 2015), thus making mandatory the assessment of the galaxy three-point correlation function or the galaxy bispectrum (see e.g. Gil-Marín et al. 2017; Slepian et al. 2017).

In order to obtain a clean detection of the BAO in the spatial distribution of galaxies (or quasars) in the two–point statistics, it is now standard to apply the concept of *reconstruction*: a technique that takes the spatial galaxy distribution back in time to a higher redshift, in which cosmic density fields are closer to linear (e.g. Eisenstein et al. 2007; Padmanabhan et al. 2012). However, a number of systematic uncertainties based on technical aspects of the observation strategy, such as the survey mask, or the radial selection function, together with other observational uncertainties with a physical background, such as galaxy bias, or redshift-space distortions, have to be taken into account in reconstruction studies. A Bayesian approach represents a natural framework to deal with these systematic uncertainties, in which a posterior distribution function (PDF) relates the linear density field to the observational data (Zaroubi et al. 1995; Kitaura & Enßlin 2008).

There are additional arguments to rely on this type of statistical approach. While mapping the linear to the non-linear density field has a clear physical foundation governed by gravity in an expanding background Universe, its inverse mapping is not trivial. The phase-space information is reduced to the spatial distribution at late cosmic times in a galaxy survey. Shell-crossing has already set in, and the trajectories of the tracers of the large-scale structure are not uniquely defined. To solve this problem, forward methods (sampling the PDF of the primordial density field, given some galaxy survey data) have been proposed in the literature (Jasche & Wandelt 2013; Wang et al. 2013, 2014; Bos, Kitaura & van de Weygaert 2019; Jasche & Lavaux 2019). See also the corresponding cosmic web analysis based on forward modelling (Nuza et al. 2014; Leclercq, Jasche & Wandelt 2015). However, these methods require to sample the initial density field in Lagrangian coordinates as a function of the final density field in the Eulerian frame. This is not only computational very expensive, but has also the drawback of adjusting the sampling procedure to the particular (i.e. particle mesh or Lagrangian perturbation theory) forward structure formation model. One of the main disadvantages of these methods is the computational cost.

To increase the computational efficiency, an effective bias prescription at the field level is used. This introduces a stochastic bias component, which requires a detailed likelihood modelling (Ata, Kitaura & Müller 2015; Mirbabayi, Schmidt & Zaldarriaga 2015; Schmidt et al. 2019). The correlation lengths between the iterations sampling the posterior distribution as reported in these papers are of the order of 1000 and producing tens of thousands of iterations, in which each time a gravity solver is applied yields only tens of independent samples. For this reason, other approaches have been proposed, in which, instead of sampling the full posterior, the maximum a posteriori is computed (Kitaura, Jasche & Metcalf 2010; Horowitz, Seljak & Aslanyan 2019; Seljak & Yu 2019). None the less, sampling the PDF has several advantages as a variety of compatible solutions can be obtained with the posterior assessment of confidence regions and therefore realistic error bars and in general, covariance matrices.

In this work, we propose an alternative approach. Grounded in the philosophy of previous works such as Monaco & Efstathiou (1999) and Kitaura (2013), our proposal goes one step further and introduces additional developments aimed at retrieving the distribution of tracers in Lagrangian space. In particular, we implement a nested Gibbs and Hamiltonian sampler, in which the final (i.e. the observed) galaxy distribution in redshift space is translated to real space at initial high-redshift coordinates defined on the light-cone. The required peculiar velocities and displacements are iteratively obtained from the primordial density field at a single high redshift with forward modelling.

Our approach is particularly efficient due to a novel higher order leapfrog algorithm applied at the solution of Hamilton equations (core of the Hamiltonian-sampling technique). We have implemented an explicit and time-reversible symplectic integrator, widely used to solve quantum field theoretical phenomena of many-body fermionic systems (see e.g. Creutz & Gocksch 1989). In a companion paper (Hernández-Sánchez et al. 2019), we show that this increases the computational efficiency by factor of about 20. Furthermore, we have implemented a non-diagonal Hamiltonian mass that includes the response operator to further increase the speed of the Hamiltonian Monte Carlo (HMC) sampler.

In order to test our method, we use a galaxy mock catalogue based on an *N*-body simulation, which reproduces clustering on the light-cone as measured from the CMASS sample (Rodríguez-Torres et al. 2016). Our results demonstrate that we can obtain unbiased linear primordial density fields up to $k \sim 0.4\, h\, \mathrm{Mpc}^{-1}$. This method promises to be especially suited for the reconstruction of BAO. Moreover, we have achieved that the reconstruction depends only on cosmological parameters, solving for dependencies on internal parameters associated with the resolution of the mesh. Another advantage of working with tracers based on the galaxy distribution is that they retain the small-scale clustering information, when taking them to higher redshifts. The displacements and velocities are obtained in Lagrangian space, while aliasing and shot-noise are accurately corrected for through a Bayesian posterior sampling. Our method uses an iterative scheme that uses only differences of particle positions and their peculiar velocities, thus being flexible to be implemented with any arbitrary gravity solver.

The method presented here shows a way of taking into account non-linear and non-local bias in the reconstruction process. This is particularly important, as the models used so far in the literature (e.g. Kitaura, Yepes & Prada 2014; Neyrinck et al. 2014; Kitaura et al. 2015) and those used within Bayesian reconstruction algorithms (e.g. Jasche & Lavaux 2019) are based on bias descriptions that have been shown to be quite inaccurate for low-mass tracers (see e.g. Pellejero-Ibañez et al. 2020). While those analytic effective Eulerian bias models can represent the distribution of luminous red galaxies (LRGs) to a good accuracy (e.g. Kitaura et al. 2016a), the approach presented in this work can be particularly relevant for surveys based on emission-line galaxies (ELGs) and bright galaxies.

Table 1. Different symbols used in the text.

Symbol	Description
$\boldsymbol{q}$	Position vector in Lagrangian coordinates
$\boldsymbol{r}$	Position vector in Eulerian coordinates (real space)
$\boldsymbol{s}$	Position vector in Eulerian coordinates (redshift space)
$\Psi(\boldsymbol{q})$	Displacement field
z_q	Redshift at the Eulerian coordinates
z_r	Redshift at the Lagrangian coordinates (real space)
z_s	Redshift at the Lagrangian coordinates (redshift space)
$\mathcal{B}$	Bias description
$\mathbf{R}$	Response function
DMDF	Dark matter density field
ALPT	Augmented Lagrangian perturbation theory
$D(z)$	Growth factor at cosmological redshift z
$\mathcal{M}$	Model

The outline of this paper is as follows. In Section 2, we discuss the theoretical framework and the main motivations of our reconstruction approach. Section 3 describes the main ingredients and operations performed within the COSMIC BIRTH (COSMological Initial Conditions from Bayesian Inference Reconstructions with THeoretical models) approach. In Section 4, we present the validation of the method. We end with conclusions.

During this work we are dealing with different estimates of redshift, viz. the cosmological redshift (in Lagrangian z_q or Eulerian coordinates z_r), and that affected by peculiar velocities z_s (in Eulerian coordinates). If nothing else indicated except z, this corresponds to z_r, i.e. the true redshift at which a galaxy resides. In Table 1, we have summarized some of the symbols used in the paper.

2 DESCRIPTION OF THE PROBLEM

In this section, we describe the theoretical context in which the COSMIC BIRTH approach resides.

2.1 The phase-space mapping problem

The data, as obtained from galaxy survey, represent a distribution of tracers (galaxies) in Eulerian redshift space. In general, there is no velocity information, except for the local Universe (see Courtois et al. 2013; Sorce et al. 2014; Tully et al. 2014). This implies on one side that an incomplete picture of the phase-space information is available, and on the other, that highly non-linear evolution (e.g. shell-crossing) is ubiquitously present.

Inferring the real-space Lagrangian coordinates $\boldsymbol{q}$ from Eulerian coordinates in redshift space, $\boldsymbol{s}$, is a complex task, as one needs already prior knowledge of the Lagrangian coordinates for the displacement field $\boldsymbol{\Psi}(\boldsymbol{q})$ and of the real-space coordinates for the peculiar velocity field along the line of sight $\boldsymbol{v}_r(\boldsymbol{q})$,[1] according to

$$\boldsymbol{q} = \boldsymbol{s} - \boldsymbol{\Psi}(\boldsymbol{q}) - \boldsymbol{v}_r(\boldsymbol{q}), \tag{1}$$

as depicted in Fig. 1. In order to illustrate the complexity of the problem, let us use the Zel'dovich approximation (Zel'dovich 1970; Shandarin & Zel'dovich 1989; White 2014). At early enough times in the evolution of perturbations in the dark matter density field (DMDF; i.e. before shell-crossing), their dynamics is well described by a potential flow with a displacement field given by $\boldsymbol{\Psi}(\boldsymbol{q}, z) = D(z)\,\nabla\Phi(\boldsymbol{q})$, where $D(z)$ is the growth factor of linear perturbations (computed throughout this work as in Heath 1977), z the cosmological redshift, and Φ is proportional to the initial gravitational potential of the perturbations satisfying Poisson's equation $\nabla^2\Phi(\boldsymbol{q}) = \delta(\boldsymbol{q})$. The position and velocity of a test particle at a fixed cosmological redshift can be written, respectively, as

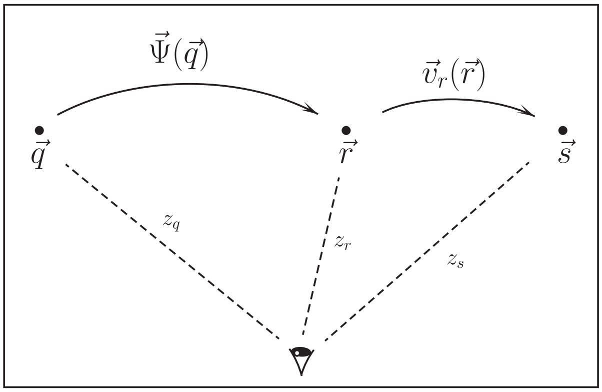

Figure 1. Sketch representing the mapping from Lagrangian coordinates $\boldsymbol{q}$ to Eulerian in redshift space $\boldsymbol{s}$. The first step consists in the mapping $\boldsymbol{q} \to \boldsymbol{r}$ mediated by the displacement field $\boldsymbol{\Psi}(\boldsymbol{q})$. The second step, mediated by the peculiar velocity field (projected along the line of sight $\boldsymbol{v}_r(\boldsymbol{r})$), takes us to $\boldsymbol{s}$. The displacement field requires prior knowledge of Lagrangian space $\boldsymbol{q}$, and the peculiar velocity field requires prior knowledge of real space $\boldsymbol{r}$.

$$\begin{aligned} \boldsymbol{r}(z) &= \boldsymbol{q} + \boldsymbol{\Psi}(\boldsymbol{q}, z) = \boldsymbol{q} + D(z)\,\nabla\Phi(\boldsymbol{q}), \\ \boldsymbol{v}(z) &= \frac{\mathrm{d}\boldsymbol{r}(z)}{\mathrm{d}\tau} = \frac{\mathrm{d}\boldsymbol{\Psi}(\boldsymbol{q}, z)}{\mathrm{d}\tau} = \dot{D}(z)\,\nabla\Phi(\boldsymbol{q}), \end{aligned} \tag{2}$$

where τ is the corresponding conformal time. The tracers defined by equation (2) occupy a three-dimensional submanifold of the entire six-dimensional phase space according to the time-dependent mapping:

$$\boldsymbol{q} \mapsto \left(\boldsymbol{q} + D(z)\,\nabla\Phi(\boldsymbol{q}),\, \dot{D}(z)\,\nabla\Phi(\boldsymbol{q})\right). \tag{3}$$

From such mapping it becomes clear that the knowledge of the initial density field $\delta(\boldsymbol{q})$ determines all posterior evolution. This is valid (without loss of generality) beyond the Zel'dovich approximation, in which case equation (2) becomes more complex (e.g. Kitaura & Hess 2013; Tassev, Zaldarriaga & Eisenstein 2013; Feng et al. 2016). The map between $\boldsymbol{r}(z)$ and $\boldsymbol{q}$ (e.g. as in equation 2) is uniquely defined (bijective) until more than one stream of dark matter exists at one spatial location (shell-crossing). Regardless of the complexity of solving the phase-space collisionless fluid equations back in time, the problem becomes irreversible having only information on the (redshift space) positions of galaxies without knowing their peculiar motions.

2.2 Previous velocity and displacement field reconstructions

In the basic reconstruction scheme widely used for BAO reconstruction (Eisenstein et al. 2007; Padmanabhan et al. 2012), the displacement and peculiar velocity fields are obtained from smoothing the galaxy field in Eulerian redshift space $\Psi(\boldsymbol{s}) = \Psi(K \otimes \delta_{\mathrm{g}}(\boldsymbol{s}))$ leading to Lagrangian coordinates expressed as

$$\boldsymbol{q} = \boldsymbol{s} - \boldsymbol{\Psi}(\boldsymbol{s}) - \boldsymbol{v}_r(\boldsymbol{s}). \tag{4}$$

This can be improved with an iterative method envisaged to effectively solve equation (1) (see Hada & Eisenstein 2019 and previous works developing this technique, e.g. Yahil et al. 1991; Monaco &

[1] Note that the velocity field is a function of the real-space Eulerian coordinates $\boldsymbol{r}$, which is in turn a function of the Lagrangian coordinates $\boldsymbol{q}$, i.e. $\boldsymbol{v}(\boldsymbol{r}) = \boldsymbol{v}(\boldsymbol{r}(\boldsymbol{q})) = \boldsymbol{v}(\boldsymbol{q})$.

Efstathiou 1999; Kitaura & Angulo 2012; Wang et al. 2012; Kitaura 2013; Kitaura et al. 2016b). Some strategies are based on seeking a unique optimal solution (e.g. Peebles 1989; Nusser & Branchini 2000; Brenier et al. 2003; Shi, Cautun & Li 2018; Sarpa et al. 2019). As mentioned in the Introduction, we are interested in forward modelling approaches within a Bayesian framework, which yield an ensemble of solutions compatible with the observations. Let us present the problem in a more formal context below leading to our chosen strategy. For an overview of other works pioneering this field we refer to the introduction given in Bos et al. (2019). We should stress here that methods like the ones explored in Monaco & Efstathiou (1999) and Hada & Eisenstein (2019) do not use a Bayesian formalism, which permits to correct for the shot noise of the discrete galaxy distribution, and the survey mask and radial selection function in the reconstruction process. A Gaussian smoothing is applied in these methods limiting the reconstruction power towards small scales. We will present here the method that permits us to deal with Lagrangian tracers within a Bayesian formalism.

3 THE COSMIC BIRTH APPROACH

The methodology of the COSMIC BIRTH approach aims at sampling the initial cosmic density field, conditional to the observed galaxy distribution on the light-cone. To this end, it relies on an iterative Gibbs-sampling method (see e.g. Kitaura & Enßlin 2008; Kitaura, Gallerani & Ferrara 2012a). The work presented here extends this approach to similarly sample displacement fields together with the peculiar velocities, as presented in Kitaura et al. (2016b). This is similar to the method presented by Kitaura (2013), with some exceptions that we will discuss below. Moreover, novel methods to sample the non-linear and non-local bias and the response function $\mathbf{R}$ in Lagrangian space are presented. The response function stands in this context for the survey geometry and radial selection function as further explained below (see equation 6). Let us discuss each step in detail below.

3.1 The Gibbs-sampling scheme

We are interested in the matter overdensity field $\delta(\boldsymbol{q})$ evaluated at Lagrangian coordinates $\boldsymbol{q}$ and defined on a regular cubical mesh with $N_{\rm c}$ cells. The data are represented by the three-dimensional galaxy distribution, as observed in Eulerian redshift space, with coordinates $\{\boldsymbol{s}^{\rm o}\}$. Given the peculiar velocities $\boldsymbol{v}$, we are in position to infer their corresponding real-space coordinates $\{\boldsymbol{r}^{\rm o}\}$. Furthermore, knowing the displacements $\boldsymbol{\Psi}$ connecting the initial cosmic times with the final ones, we can compute their corresponding Lagrangian coordinates $\{\boldsymbol{q}^{\rm o}\}$, as discussed in the previous section. The galaxy number counts on the mesh $\boldsymbol{N}_{\rm g}$ can be related to the matter density field according to our likelihood model and a bias description $\mathcal{B}$ presented in Section 3.5. Furthermore, we use a response function $\mathbf{R}$, which accounts for the survey geometry, or angular completeness, and for the radial selection function. In particular, the joint PDF of all the above-mentioned variables can be sampled within a Gibbs-sampling scheme based on the corresponding conditional PDFs:

$$\begin{aligned}
\delta(\boldsymbol{q}) &\curvearrowleft \mathcal{P}_\delta\left(\delta_{\boldsymbol{q}}|\{\boldsymbol{q}^{\rm o}\}, \mathbf{R}_{\boldsymbol{q}}, \mathbf{C}_{\boldsymbol{q}}\left(\{p_{\rm c}\}\right), \{\mathcal{B}_{\boldsymbol{q}}\}\right), \\
\{\boldsymbol{r}^{\rm o}\} &\curvearrowleft \mathcal{P}_r\left(\{\boldsymbol{r}^{\rm o}\}|\{\boldsymbol{s}^{\rm o}\}, \{\boldsymbol{v}^z\left(\delta_{\boldsymbol{q}}, f_\Omega^z\right)\}, \mathcal{M}_v\right), \\
\{\boldsymbol{q}^{\rm o}\} &\curvearrowleft \mathcal{P}_q\left(\{\boldsymbol{q}^{\rm o}\}|\{\boldsymbol{r}^{\rm o}\}, \{\boldsymbol{\Psi}_{\boldsymbol{q}}^z\}, \mathcal{M}_\Psi\right), \\
\mathbf{R}(\boldsymbol{q}) &\curvearrowleft \mathcal{P}_R\left(\mathbf{R}_{\boldsymbol{q}}|\mathbf{R}_{\boldsymbol{s}}, \{\boldsymbol{\Psi}_{\boldsymbol{q}}^z\}, \mathcal{M}_\Psi\right), \\
\{\mathcal{B}(\boldsymbol{q})\} &\curvearrowleft \mathcal{P}_B\left(\{\mathcal{B}_{\boldsymbol{q}}\}|\{\mathcal{B}_{\boldsymbol{s}}\}, \{\boldsymbol{\Psi}_{\boldsymbol{q}}^z\}, \mathcal{M}_\Psi\right),
\end{aligned} \tag{5}$$

where the subscripts q and s stand for Lagrangian real-space and Eulerian redshift-space coordinates, respectively. The curved left arrows stand for the sampling process. The superscript z stands for redshift bin. $\mathcal{M}_v$ and $\mathcal{M}_\Psi$ represent the models describing peculiar motions and displacement fields, respectively. The growth rate f_Ω^z will be further discussed in Section 3.3.

The approach described above has the great advantage of being general for any structure formation model, as only the initial and final positions with their peculiar motions are needed. In this study, we only consider the approach provided by the augmented Lagrangian perturbation theory (ALPT; Kitaura & Hess 2013), which is being successful describing clustering down to a few Mpc scales and has been implemented for the generation of halo mock catalogues generation based on bias mapping methods (Kitaura et al. 2016b; Balaguera-Antolínez et al. 2019).

Figs 2 and 3 depict, with a flowchart, the main steps followed within the Gibbs-sampling method. In the following subsections we summarize the first three main crucial Gibbs-sampling steps and the corresponding assumptions (additional steps will be subsequently presented).

3.2 Step 1: sampling the linear density field

The continuous primordial dark matter field is sampled assuming that the observed galaxies are identified in real Lagrangian space at high redshift. This is done within a Bayesian framework, in which a lognormal-Poisson PDF is assumed. The lognormal PDF stands for the prior of the dark matter distribution, while the Poisson PDF represents the likelihood describing the distribution of discrete Lagrangian space tracers (e.g. Kitaura et al. 2010). We note that the lognormal prior assumes on one side a comoving Lagrangian framework, and, on the other, that tracers can be uniquely followed, i.e. neglecting shell-crossing (Coles & Jones 1991; Kitaura & Angulo 2012). This precisely applies for Lagrangian tracers at high redshift. Also, the logarithmic transformation of the normalized density log (1 + δ) (with $\delta = \rho/\bar{\rho} - 1$) tends towards the overdensity field for $|\delta| \ll 1$.

On the other hand, Poissonity is a reasonable assumption for homogeneously distributed tracers at high redshift, e.g. before gravity introduces small-scale clustering. Such small-scale clustering generates over-Poisson dispersions at the scale of the subvolume element at which the galaxy number counts are defined, since at larger scales an inhomogeneous Poisson distribution accounts for the large-scale clustering modulated by the DMDF (e.g. Peebles 1980; Saslaw 1989; Sheth 1998; Kitaura et al. 2014; Neyrinck et al. 2014; Ahn et al. 2015). In any case such a deviation from Poissonity can also be included in a Bayesian framework (Ata et al. 2015).

The density estimation step is the bottleneck in our computations. We would like to stress that COSMIC BIRTH achieves a high efficiency with the development of a novel version of the HMC sampling introduced by Jasche & Kitaura (2010), and studied in detail in a companion paper (Hernández-Sánchez et al. 2019). For further reference to the basic implementation see also Ata et al. (2017). In particular, we have introduced, as a novel ingredient, a higher order discretization of Hamilton equations using subsequent second-order Leapfrog operators. In this work, we use the fourth-order discretization, which consists of a forward time step integration, followed by a backward one, with a third and final forward one with different time step lengths following Creutz & Gocksch (1989).

In addition, we introduce a strategy to deal with non-diagonal Hamiltonian mass matrices including the survey geometry, which

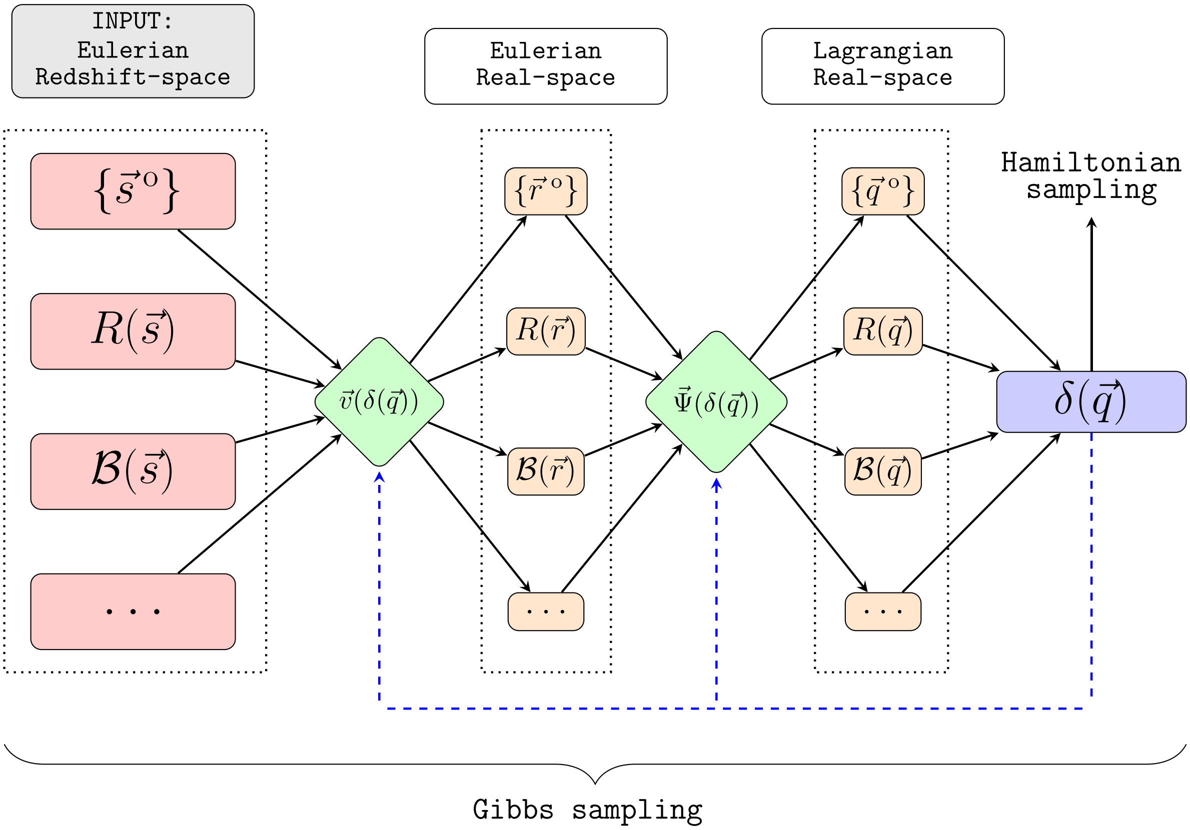

Figure 2. Flowchart depicting the Gibbs-sampling scheme of BIRTH to sample the density field $\delta(\boldsymbol{q})$ at initial cosmic times. The input data reside in Eulerian redshift space and can be transformed to real space with a given peculiar velocity field $\boldsymbol{v}$. Once the data have been transformed to Lagrangian space with the displacement field $\boldsymbol{\Psi}$, we can sample the corresponding density field with higher order Hamiltonian-sampling. We note that this scheme can be extended to account for more variables of the model (power spectrum, growth rate, etc.).

acts as a *pre-conditioner* to additionally speed up the algorithm (see Appendix B).

The Hamiltonian sampler could have assigned a larger number of tracers in Lagrangian space, as available in the Eulerian light-cone, i.e. the observed galaxy distribution. This was done in Kitaura et al. (2012b), Heß, Kitaura & Gottlöber (2013), and Kitaura (2013). Here we want to control the evolving bias on the light-cone (see Section 3.5). Therefore we will restrict in this first paper the Lagrangian tracers to be equal in number to the Eulerian ones. We note, however, that the bias does not change, if the number of Lagrangian tracers is equal for each Eulerian tracer, as the overdensity field does not change. In case, one would consider a different number of tracers depending for instance on the location of the galaxy in the cosmic web this picture becomes more complicated, although not unsolvable. We leave this investigation for future work.

3.3 Steps 2 and 3: displacements and peculiar velocity fields

The Lagrangian tracers are sampled assuming that the initial cosmic density field is known. This can be achieved using a given structure formation model, yielding the displacement and peculiar velocity fields at different redshifts from that initial field. These in turn can be used to obtain the Lagrangian tracers that correspond to the observed tracers in Eulerian space. We note that this step is similar to the analysis of N-body simulations for which the initial conditions are known and the particles composing a halo are traced back to high redshift (e.g. Ludlow & Porciani 2011). The additional uncertainty in our study comes from the lower resolution at which we reconstruct the initial Gaussian field, which is in general not high enough to resolve the haloes hosting the observed galaxies. Also, galaxy bias and redshift-space distortions contribute to this uncertainty, as we will discuss below.

In this second step COSMIC BIRTH obtains the real-space position for each galaxy, given its observed redshift-space $s^{\rm obs}$ position (required for the first step). The latter is obtained by sampling the peculiar velocities $\{\boldsymbol{v}(\boldsymbol{\delta}, f_\Omega)\}$ (with the growth rate given by $f_\Omega \equiv {\rm d}\log D(a)/{\rm d}\log a$), assuming that the density field and the growth rate f_Ω are known.[2] One can add a dispersion term to the displacement and peculiar velocity field accounting for the uncertainty. In practice this is assumed to be Gaussian distributed, and set to low standard deviations of about 1 h^{-1} Mpc (see Kitaura et al. 2016b; Ata et al. 2017).

In a further publication we will explore in more detail the redshift-space distortions corrections achieved with COSMIC BIRTH. We should also note that the solution to the Lagrangian-to-Eulerian mapping problem is not solved here in the same way to the approach presented in Kitaura (2013). While in the latter approach a large number of large-scale structure tracers are displaced forward in time and then linked to observed galaxy distribution (in a likelihood comparison

[2] We note that assuming a wrong growth rate will yield an anisotropic reconstructed density field. This was recently investigated (Granett et al. 2015) by jointly sampling the anisotropic power spectrum including the growth rate and the redshift-space density field.

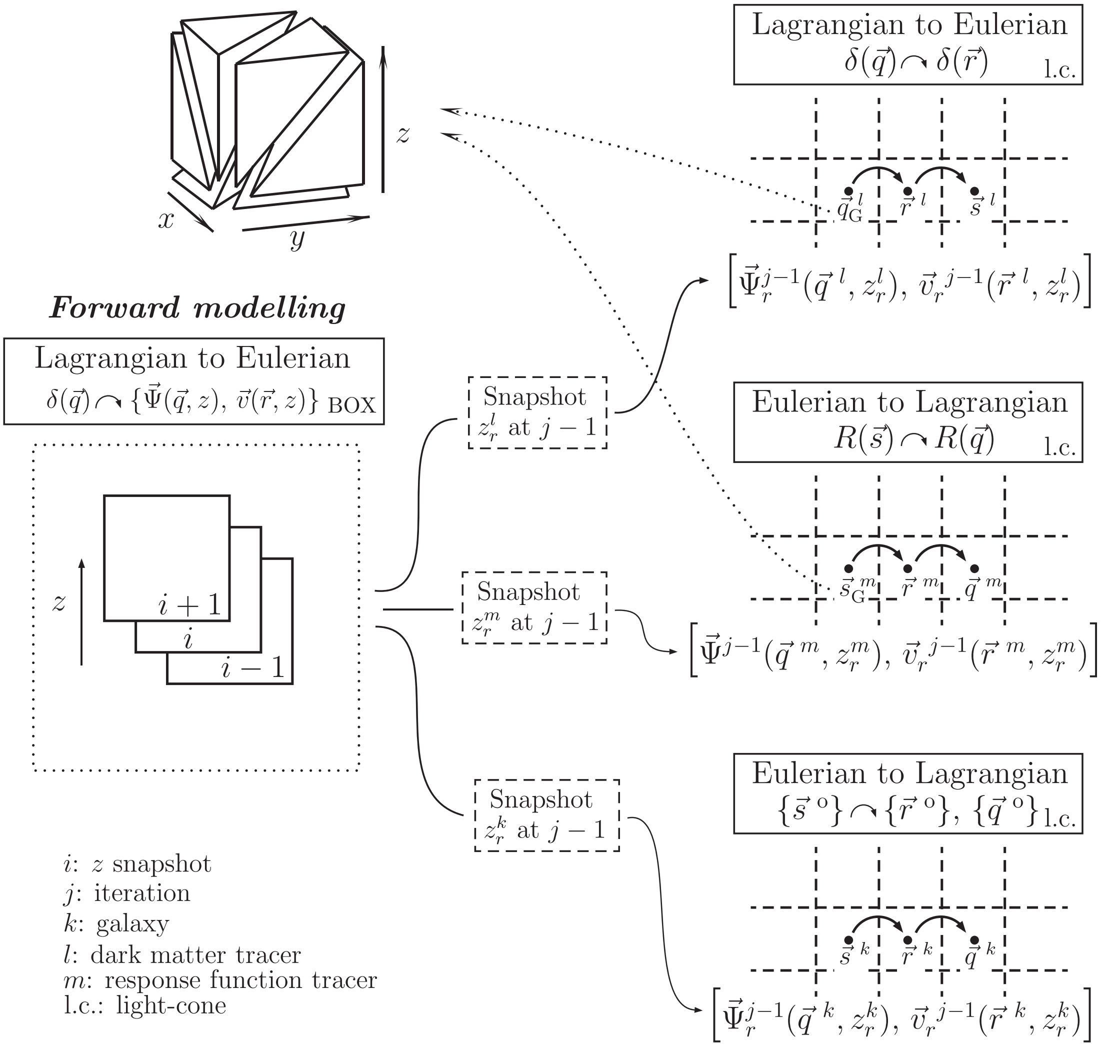
Lagrangian to Eulerian
$\delta(\vec{q}) \curvearrowright \delta(\vec{r})$ l.c.
z
x
y
$\vec{q}_{\rm G}^{\,l}$ $\vec{r}^{\,l}$ $\vec{s}^{\,l}$
$\left[\vec{\Psi}_r^{j-1}(\vec{q}^{\,l}, z_r^l),\ \vec{v}_r^{\,j-1}(\vec{r}^{\,l}, z_r^l)\right]$
Forward modelling
Lagrangian to Eulerian
$\delta(\vec{q}) \curvearrowright \{\vec{\Psi}(\vec{q}, z),\ \vec{v}(\vec{r}, z)\}$ BOX
Snapshot z_r^l at $j-1$
Eulerian to Lagrangian
$R(\vec{s}) \curvearrowright R(\vec{q})$ l.c.
z
$i+1$
i
$i-1$
Snapshot z_r^m at $j-1$
$\vec{s}_{\rm G}^{\,m}$ $\vec{r}^{\,m}$ $\vec{q}^{\,m}$
$\left[\vec{\Psi}^{j-1}(\vec{q}^{\,m}, z_r^m),\ \vec{v}_r^{\,j-1}(\vec{r}^{\,m}, z_r^m)\right]$
Eulerian to Lagrangian
$\{\vec{s}^{\,o}\} \curvearrowright \{\vec{r}^{\,o}\},\ \{\vec{q}^{\,o}\}$ l.c.
Snapshot z_r^k at $j-1$
i: z snapshot
j: iteration
k: galaxy
l: dark matter tracer
m: response function tracer
l.c.: light-cone
$\vec{s}^{\,k}$ $\vec{r}^{\,k}$ $\vec{q}^{\,k}$
$\left[\vec{\Psi}_r^{j-1}(\vec{q}^{\,k}, z_r^k),\ \vec{v}_r^{\,j-1}(\vec{r}^{\,k}, z_r^k)\right]$

Figure 3. This sketch illustrates the various Lagrangian to Eulerian and vice versa mappings performed within COSMIC BIRTH. At the bottom right it is illustrated how galaxies are mapped to real Lagrangian space looking up the information from a forward simulation based on the previous iteration yielding outputs at different redshift snapshots $(\dots, i-1, i, i+1, \dots)$, as shown on the left. The light-cone dark matter field calculation from Lagrangian to Eulerian space and the response function mapping from Eulerian to Lagrangian space happen in the same loop, each of one looking up the corresponding redshift snapshot information, when going through the cells of the mesh. For both these cases a Lagrangian tetrahedral tessellation is applied. To this end the corresponding tracers are defined at the cell centre and thus denoted with a subscript G.

step), here we solve equation (1) by evaluating the forward computed displacement and velocity fields at the Lagrangian locations of the previous iteration (see Fig. 3). This is not yielding a completely self-consistent relation between the sampled initial Gaussian field and the final positions of tracers, as the Gaussian fields change from iteration to iteration given various uncertainties modelled in the Bayesian framework. Therefore, we expect some improvement at least on small scales, if the approach formulated in Kitaura (2013) is implemented in the COSMIC BIRTH code. As anticipated, we will investigate this in detail in future work.

3.4 Step 4: response function in Lagrangian space

The problem arising from the type of approach described in this section is that the observables are obtained in Eulerian space, while the reconstruction of the matter density field is performed in Lagrangian space, under the assumption that the data are in that space. This is still an unsolved problem for the survey geometry or the radial selection function. Previous studies (Kitaura 2013) have augmented empty regions with some mock galaxies and inverse weighted the selection function to mitigate these issues. Since this operation is particularly dangerous when the selection function acquires very low values (see discussion in Kitaura et al. 2009), such studies were restricted only to small cosmological volumes (Kitaura et al. 2012b).

One can improve this in a number of ways. The key concept is data augmentation to enable a balanced likelihood analysis (see e.g. Tanner 1993). We enumerate several strategies as follows.

(i) One possibility is based on the production of data exactly compensating for the incompleteness of the survey in each Gibbs-sampling iteration. This can be done according to some bias model and the density field obtained in a given iteration. Marginalization over such augmented data could be done by sampling, in each Gibbs-sampling iteration, new augmented data, discarding the previous ones (the true observed data are untouched during the iterations). Note that this approach needs an accurate Eulerian bias model, which is particularly difficult to achieve (Pellejero-Ibañez et al. 2020).

(ii) A second approach can be that of introducing a noise component (as is done with Wiener filtering; e.g. Zaroubi et al. 1995; Horowitz et al. 2019), which depends on the completeness, being larger in less observed areas. The inconvenience of this approach is that in our particular case one would need some degree of arbitrary fine-tuning to get sensible results as the level of noise is a complex function of the completeness for which we lack a proper model (e.g. Zaroubi, Hoffman & Dekel 1999).

(iii) The most natural option consists of including the completeness in a response function and let the Bayesian model compensate the survey mask and radial selection function in the reconstructed initial cosmic density field (see Zaroubi et al. 1995; Kitaura & Enßlin 2008). This has been done for the first time connecting Lagrangian to Eulerian space including cosmic evolution in Jasche & Wandelt (2013), and later adopted by Wang et al. (2013) and Bos et al. (2019). All these approaches compute gradients of structure formation demanded within the Hamiltonian-sampling. Also, an accurate description of Eulerian bias is requested, without which those methods are likely to downweight the data (see e.g. Jasche & Lavaux 2019, where the likelihood is downweighted with a factor of 0.3).

(iv) We propose here to calculate the response function in Lagrangian space to be able to apply the standard Bayesian approach. We explain this in detail below.

The response function can be straightforwardly computed for the radial selection function, as we just have to calculate it based on the reconstructed galaxy sample at Lagrangian coordinates. However, the angular survey mask is not trivial to compute in Lagrangian space. First we have to project it to the three-dimensional space as it is introduced in Kitaura et al. (2009). Then that Eulerian field has to be mapped to Lagrangian space using the reconstructed forward displacement field on the light-cone. In particular, we assume that each cell centre $\boldsymbol{r}_{\rm g}$ represents a response function tracer. We need to keep track of the Lagrangian coordinates of each cell centre in the same way we do it with the galaxies. This enables us in principle to make a mapping of the response function to Lagrangian space. However, the finite number of tracers (finite resolution of the grid) yields inaccurate estimates of this mapping.

Therefore we resort to Lagrangian tetrahedral tessellation (see Abel, Hahn & Kaehler 2012; Shandarin, Habib & Heitmann 2012; Hahn, Abel & Kaehler 2013), which makes a tessellation of the mesh into tetrahedrons and uses the positions and displacement field information to get accurate density estimates even on coarse resolutions (e.g. Balaguera-Antolínez et al. 2019). See also the works by Falck, Neyrinck & Szalay (2012) and Neyrinck (2012, 2013).

Let us call the resulting three-dimensional projected angular mask as $\boldsymbol{w}_\alpha$, and refer to the corresponding the angular response function as

$$\mathbf{R}_\alpha = \boldsymbol{w}_\alpha \mathbb{1}. \tag{6}$$

We note that the angular completeness does not care about real or redshift space, which affect only the radial direction. We consider thus the redshift z_s as the final one of the displacement field (in the Zel'dovich approximation this would be $\Psi(q, z) = D(z_r)\Psi(q)$).

The radial selection function is computed from the z-distribution of large-scale structure tracers normalized by the volume enclosed in shells (or divided by z^2 as it is described in Ata et al. 2017). We do this computation in Lagrangian space, as we need all quantities defined in that space in order to perform the Step 1. Let us refer to the radial selection function part of the response function as

$$\mathbf{R}_r = \boldsymbol{w}_r \mathbb{1}, \tag{7}$$

where $\boldsymbol{w}_r$ is the three-dimensional projected spherical symmetric radial selection function. The total response function $\mathbf{R}$ will be the product of $\mathbf{R}_\alpha$ and $\mathbf{R}_r$:

$$\mathbf{R} = \mathbf{R}_\alpha \cdot \mathbf{R}_r. \tag{8}$$

We note that (iteratively) computing the radial selection function in real (Lagrangian) space prevents the so-called 'Kaiser rocket' effect (Kaiser 1987; Nusser, Davis & Branchini 2014). The Lagrangian framework we are using has another advantage. We can define the radial selection function as described above ignoring light-cone effects, as we are performing the reconstruction at a single snapshot at high redshift. However, direct reconstructions of the DMDF in Eulerian space require to take into account a cosmological selection function (see Granett et al. 2015).

3.5 Step 5: sampling the galaxy–dark matter bias relation

The description of the galaxy distribution (in our context: galaxy number counts per cell mapping the observed volume on to a mesh) with respect to the large-scale dark matter field (defined on the same mesh) requires effective bias models, encoding the underlying physics of galaxy formation in a non-linear, non-local functional dependence. The large-scale bias can be measured in redshift bins (and galaxy populations according to various properties) using different probes of clustering (e.g. Verde et al. 2002; Conway et al. 2005; Seljak et al. 2005; Cresswell & Percival 2009; Lindsay et al. 2014; Gil-Marín et al. 2015; Balaguera-Antolínez et al. 2018; Pan et al. 2020). The galaxy bias is in general a non-linear function of the underlying continuous dark matter field. In the attempt of modelling such a relation, a Taylor expansion has been suggested both as a function of the dark matter (i) overdensity field (Fry & Gaztanaga 1993), and (ii) to its logarithm (Cen & Ostriker 1992). In fact the latter expansion corresponds, truncated to first order, to a power law, giving already a fair description at the two-point statistics (see e.g. de la Torre & Peacock 2013). However, it has been shown that a threshold bias based on the peak split-background picture (e.g. Kaiser 1984) is crucial for an accurate description of the three-point statistics (see e.g. Kitaura et al. 2014, 2015). This model has been refined to have a smoother drop-off behaviour towards the low-density regime by Neyrinck et al. (2014) and has been successfully applied to reproduce the LRG distribution of the Baryon Oscillation Spectroscopic Survey (BOSS; Kitaura et al. 2016a). The expected galaxy number counts is then given by

$$\rho_{\rm g}(\boldsymbol{r}, z) = \gamma(z)\mathcal{B}(\boldsymbol{r}, z), \tag{9}$$

with a normalization of

$$\gamma(z) = \frac{\bar{N}(z)}{\langle \mathcal{B}(\boldsymbol{r}, z)\rangle}, \tag{10}$$

and a non-linear deterministic bias given by

$$\mathcal{B}(\boldsymbol{r}, z) = \exp\left[-\left(\frac{\rho(\boldsymbol{r}, z)}{\bar{\rho}(z) b_\rho(z)}\right)^{b_\epsilon(z)}\right]\left(\frac{\rho(\boldsymbol{r}, z)}{\bar{\rho}(z)}\right)^{b_{\rm p}(z)}, \tag{11}$$

where the density is linked to the overdensity field through $\rho(\boldsymbol{r}, z) = \bar{\rho}(z)(1 + \delta(\boldsymbol{r}, z))$. However, this simple non-linear model lacks a proper non-local bias description (see e.g. McDonald & Roy 2009), which can be modelled through the tidal field to second order (see e.g. Balaguera-Antolínez et al. 2019, 2020). A complete bias description

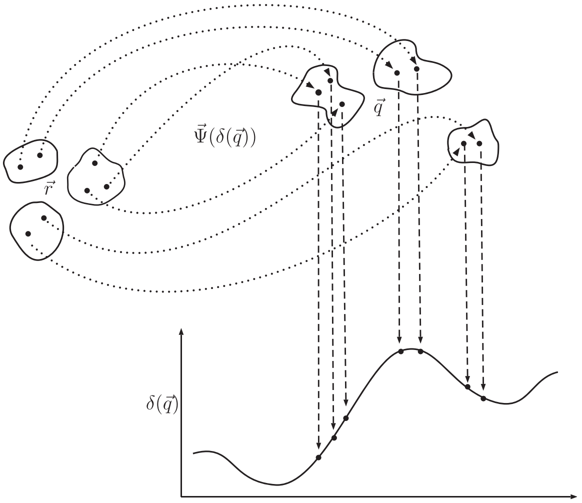

Figure 4. This sketch illustrates the relation between galaxy and halo bias at Eulerian and Lagrangian space. Galaxies (represented by dots) are depicted as tracers of haloes (represented by asymmetric regions). The corresponding protohaloes are non-spherical and do not necessarily trace the peaks of the initial cosmic density field. This is accounted for in COSMIC BIRTH when mapping the galaxies to Lagrangian space and using a bias without selecting the peaks, i.e. without threshold bias. A higher galaxy number density yields a more accurate description of the protohalo regions. This hints towards the advantage of using multiple galaxies at the same Eulerian position instead of varying the mass, as these will be mapped differently in Lagrangian space.

also demands in principle the dependence with the initial cosmic field (e.g. Desjacques, Jeong & Schmidt 2018) and there are current attempts to include this within a Bayesian context.

The goal of this work is to find a practical Lagrangian bias description that can be directly derived from the observations, assuming that the tracers of the large-scale structure reside in Lagrangian space. This is achieved within our Gibbs-sampling scheme through an Eulerian to Lagrangian mapping that already accounts for non-local and non-linear bias, simplifying the bias relation in Lagrangian space. This is represented in a sketch in Fig. 4.

In the remainder of this section we will derive a complete formalism that connects the observed redshift large-scale clustering over the large-scale Lagrangian bias, to a non-linear Lagrangian bias model including the dependence on the chosen mesh resolution to represent the galaxy number counts and the dark matter field.

3.5.1 *Eulerian large-scale bias*

The clustering of galaxies in redshift space with respect to some fiducial cosmology provides a measure of the large-scale bias. Following Ata et al. (2017), given a redshift z one can define the ratio between the galaxy correlation function in redshift space at z ($\xi_{\rm g}^s(z)$) and the matter correlation function in real space at z ($\xi_{\rm M}(z)$) as

$$b^s(z) \equiv \sqrt{\left.\frac{\xi_{\rm g}^s(z)}{\xi_{\rm M}(z)}\right|_{\rm LS}}. \qquad (12)$$

The quantity $\xi_{\rm g}^s(z)$ can be obtained from the data without having to assume any information of bias or growth rate. Furthermore, one can use the Kaiser factor $K = 1 + (2/3)f_\Omega/b + (1/5)(f_\Omega/b)^2$ (where f_Ω denotes the growth rate; Kaiser 1987) to relate the galaxy correlation function in redshift space to the matter correlation function in real space, $\xi_{\rm g}^s(z) = K(z)\,b^2(z)\,\xi(z)$. Combining these expressions we find a quadratic expression for $b(z)$ for each redshift z, with a positive solution given by (see e.g. Ata et al. 2017)

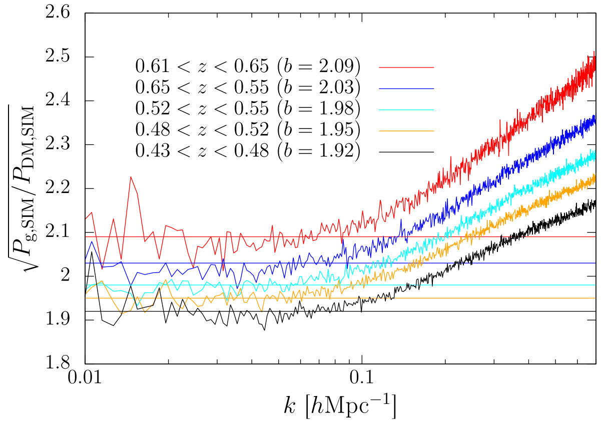

Figure 5. Large-scale bias obtained from the power spectrum ratio between the galaxy and the dark matter field at different redshift snapshots from the BigMD simulation [subhalo abundance matching (SHAM) catalogues and dark matter particles, respectively]. For illustrative purposes five bins are shown from the 10 to 20 used in our study.

$$b(z) = -\frac{1}{3}f_\Omega(z) + \sqrt{-\frac{4}{45}f_\Omega(z)^2 + (b^s(z))^2}. \qquad (13)$$

In our case study using the light-cone mock galaxy catalogue for CMASS galaxies we find a bias as a function of redshift as illustrated in Fig. 5.

3.5.2 *Lagrangian large-scale bias*

Once we have the Eulerian large-scale bias given by equation (13), we can translate it to any higher redshift. To that aim, we assume that that the bias from equation (13) can be expressed as $\delta_{\rm g}(z) = b(z)\delta(z)$. On top of this, given (i) the conservation in the number of galaxies, and (ii) that the galaxies follow the same velocity field as the underlying DMDF, one can demonstrate that the large-scale bias can be written at any other redshift in terms of the linear growth factor $D(z)$ as (see e.g. Nusser & Davis 1994; Fry 1996; Percival & Schäfer 2008)

$$b(z_q) = (b(z) - 1)\,\frac{D(z)}{D(z_q)} + 1. \qquad (14)$$

We note that we do not need to include stochasticity in this relation, as introduced by Tegmark & Peebles (1998), since we are not trying to model the different galaxy populations at different redshifts including galaxy formation, but the large-scale bias evolution of a given galaxy population. In fact this has been studied in detail in Birkin et al. (2019).

3.5.3 *Lagrangian non-linear bias*

Hitherto, we have only considered the bias at large scales (i.e. in the limit of $k \rightarrow 0$). If we aim at describing the DMDF on a mesh of Mpc scales resolution, we need to use a non-linear description of DMDF. In fact, a typical bias of $\sim$2 (e.g. for LRGs) is translated through equation (14) to a bias of about 60 at $z = 100$. If our cell resolution is high enough of producing overdensities larger than $|\delta| > 10^{-2}$, this implies that a linear model would yield negative densities, i.e. $\delta < -1$.

One of the simplest models we can assume is a power-law bias:

$$\rho_{\rm g}(\boldsymbol{q}) = \gamma(z_q)(1+\delta(\boldsymbol{q}))^{b(z_q)\, f_{\rm b}(z_q)}, \tag{15}$$

where $\gamma(z_q)$ is a normalization constant and $f_{\rm b}$ is a correction factor that ensures a correct large-scale bias (see e.g. Ata et al. 2017). This model does not include threshold bias (see equation 11) inherent to the peak background-split model (see e.g. Kaiser 1984; Schmidt, Jeong & Desjacques 2013, and references therein) . This is consistent with the picture of the protohaloes associated with haloes after cosmic evolution, which are not tracers of the peaks of the initial cosmic density field, but can be tracing the whole density regime (see e.g. Ludlow & Porciani 2011). In our framework, represented in Fig. 4, galaxies tracing haloes are mapped to Lagrangian space tracing the protohaloes in the entire density field. In a natural way the resulting protohalo regions are not spherical symmetric already effectively ensuring a non-local mapping in Eulerian space (see e.g. Sheth, Chan & Scoccimarro 2013). The framework presented here allows to be extended to account for complex Lagrangian bias components, if that would be required. We note, however, that recent works do not find important non-local bias contributions in Lagrangian space, except for very massive haloes (Modi, Castorina & Seljak 2017; Abidi & Baldauf 2018). This implies that as long as the Lagrangian tracers used to reconstruct the density field are not very massive, we can neglect additional non-local bias terms (see Fig. 4).

It is important to note that the normalization in equation (15) depends on the non-linear bias model, and is only equal to the galaxy number density $\bar{N}$ for bias unity:

$$\gamma(z_q) = \frac{\bar{N}}{\langle(1+\delta(\boldsymbol{q}))^{b(z_q)\, f_{\rm b}(z_q)}\rangle}. \tag{16}$$

The problem associated with the model represented by equations (15) and (16) is its dependency on the mesh resolution on which $\rho_{\rm g}(\boldsymbol{q})$ and $\delta(\boldsymbol{q})$ are defined, via the factor $f_{\rm b}$, and thereby on input parameters used for the representation of the data in our code. To circumvent this situation, we obtain a connection between the power-law bias of equation (16) (specifically, the parameter $f_{\rm b}$) with the large-scale bias as predicted by the renormalized perturbation theory (RPT; see e.g. McDonald & Roy 2009; Desjacques et al. 2018). We present the derivation in Appendix A.

Given the lack of a solid analytical framework that predicts the non-linear bias, we propose here to derive it numerically. This ansatz is inspired by RPT (see Appendix A), which encodes the non-linear dependence on the resolution in the variance of the field. We can define an effective power-law bias by $b_{\rm eff}(z) = b(z)\, f_{\rm b}(z)$. The large-scale bias can be obtained from the ratio of the galaxy to dark matter overdensity variances:

$$b(z) \equiv \sqrt{\frac{\sigma^2_{K_{\rm g}}(z)}{\sigma^2_K(z)}}, \tag{17}$$

with the variances given by

$$\sigma^2_K(z) = \langle(K \circ \delta(\boldsymbol{q}, z))^2\rangle \tag{18}$$

and

$$\sigma^2_{K_{\rm g}}(z) = \langle(K \circ \delta_{\rm g}(\boldsymbol{q}, z)[b_{\rm eff}])^2\rangle \tag{19}$$

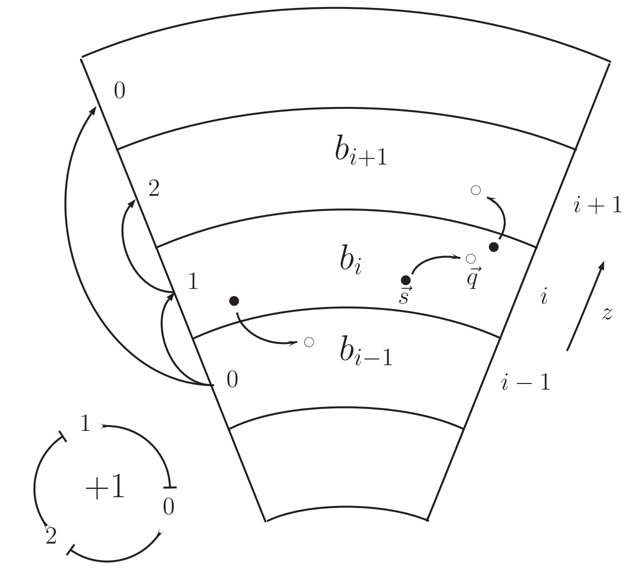

Figure 6. This is a sketch of the light-cone of a galaxy survey divided in redshift bins $\dots, i-1, i, i+1, \dots$, with the corresponding large-scale bias $\dots, b_{i-1}, b_i, b_{i+1}, \dots$. A galaxy may stay in its redshift bin when doing reconstruction $\boldsymbol{s} \rightarrow \boldsymbol{q}$, jump to a lower, or to a higher redshift bin. COSMIC BIRTH goes through the data in a cyclic order of even permutations: ...-0-1-2-0-..., requiring the storage in RAM of the displacements and velocities of only three redshift snapshots at once.

for the dark matter and the galaxy field, respectively, and K being a Gaussian kernel with a smoothing scale of 50–100 h^{-1} Mpc[3] The model for the galaxy overdensity is accordingly written as

$$\delta_{\rm g}(\boldsymbol{q}, z)[b_{\rm eff}] \equiv \bar{N}(z)\,\frac{(1+\delta(\boldsymbol{q}, z))^{b_{\rm eff}}}{\langle(1+\delta(\boldsymbol{q}, z))^{b_{\rm eff}}\rangle} - 1. \tag{20}$$

Now we have all the ingredients to obtain the non-linear bias correction factor $f_{\rm b}(z)$, which ensures that the large-scale bias (equation 17) is recovered for large smoothing radii from equation (20). We do this iteratively using a Newton–Raphson method.[4]

3.5.4 *Bias mixing between redshift bins*

So far we have found a Lagrangian bias description based on some clustering measurements in Eulerian redshift space. Since those are naturally done in redshift bins, we end up having the Lagrangian bias defined on shells in redshift distance. This assumes that the Eulerian to Lagrangian mapping of tracers keeps spherical shells, however those are distorted in the same (reverse) way as BAO spheres are distorted through cosmic evolution. If the bias is interpolated to obtain a smooth varying function in redshift, then the changes in redshift from z_s to z_q according to Fig. 1 are not large and it can be assumed that the Lagrangian bias of a galaxy is the same evaluated at both distances. If one decides to keep a binned bias, then a galaxy in one bin might jump to a higher or lower redshift bin as shown in Fig. 6. In such scenario one would need to associate with galaxies jumping to a lower (higher) redshift bin the bias from their original

[3] After verifying that the results do not change for our volume using different smoothing scales, we chose a scale of 50 for our numerical tests.
[4] We choose in our calculations an accuracy of eps $= 10^{-5}$, which typically is achieved after 3–5 iterations for each population of galaxies and for each redshift bin.

higher (lower) one. This has been actually implemented in COSMIC BIRTH and we show results for both cases below.

4 VERIFICATION OF THE COSMIC BIRTH CODE

In this section, we will describe the data used to verify the COSMIC BIRTH and present and discuss the results after running the COSMIC BIRTH code on them.

4.1 Data used in this work

To validate the reconstruction method presented in this paper, we use the Data Release 12 (DR12) of the BOSS (Dawson et al. 2013). The BOSS survey uses the Sloan Digital Sky Survey (SDSS) 2.5-m telescope at Apache Point Observatory (Gunn et al. 2006) and the spectra are obtained using the double-armed BOSS spectrograph (Smee et al. 2013). The data are then reduced using the algorithms described in Bolton et al. (2012). The target selection of the CMASS and LOWZ samples, together with the algorithms used to create large-scale structure catalogues (the MKSAMPLE code), is presented in Reid et al. (2016).

We restrict this analysis to the CMASS sample of LRGs, which is a complete sample, nearly constant in mass and volume, limited between the redshifts $0.43 \le z \le 0.7$ (see Anderson et al. 2014, for details of the targeting strategy). We use the N-body-based mock galaxy catalogue constructed to match the clustering bias, survey mask, selection functions, and number densities of the BOSS DR12 CMASS galaxies on the light-cone.

The mock galaxy catalogue used in this study was presented in Rodríguez-Torres et al. (2016) and was extracted from the BigMDPL N-body simulation,[5] one of the MultiDark simulation project, which was performed using the GADGET-2 code (Springel 2005). The BigMDPL was run with 3840^3 particles on a volume of $(2.5\,h^{-1}\,\mathrm{Gpc})^3$ assuming Λ cold dark matter (ΛCDM) *Planck* cosmology with $\{\Omega_\Lambda = 0.6928, \Omega_\mathrm{M} = 0.307, \Omega_\mathrm{b} = 0.0482, \sigma_8 = 0.828, n_\mathrm{s} = 0.961\}$, and a Hubble constant ($H_0 = 100\,h\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$) given by $h = 0.677$. Haloes and subhaloes were identified using the ROCKSTAR halo finder (Behroozi, Wechsler & Wu 2013). The DMDF on the light-cone has been constructed with the redshift snapshots between $z = 0.43$ and $z = 0.7$ using the stored data from the BigMDPL simulation, i.e. 0.5 per cent of the particles. As a further preparation of the data, we computed the response function following the description in Section 3.4. In particular, the angular mask was calculated using the MANGLE software package (Hamilton & Tegmark 2004; Swanson et al. 2008). For the time being we will assume the power spectrum to be known with the exact cosmology (i.e. used to construct the mock galaxy catalogue). We note that the large-scale bias on redshift bins comes as an input computed as shown in Fig. 4.

4.2 Results

We consider in our analysis cubical volumes of $L = 3200\ h^{-1}$ Mpc side length with 256^3 cells, i.e. a cell resolution of $12.5\ h^{-1}$ Mpc.[6] This setting is identical to the one in Ata et al. (2017, and we refer the reader to this work for further details). We have performed a series of runs with a variation of settings.

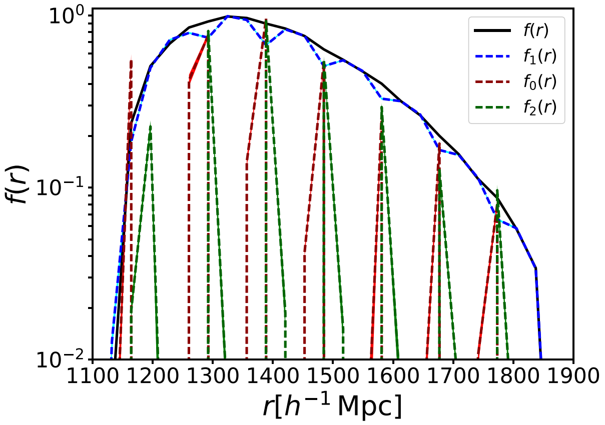

Figure 7. Radial selection function: original including all galaxies in Eulerian space (black solid line), after reconstruction in Lagrangian space corresponding to galaxies (i) staying in their Eulerian redshift bin (subscript 1 and dashed blue line), (ii) jumping to a lower redshift (subscript 0 and dashed red line), and (iii) jumping to a higher redshift bin (subscript 2 and dashed green line).

From now on the reference calculation is dubbed run A. This run includes 20 redshift snapshots in the range $0.35 < z < 0.8$, including Lagrangian tetrahedral tessellation in the angular response function transformation to Lagrangian coordinates (see Section 3.4), and bias interpolation (see Section 3.5.4). We note that we need an enlarged redshift range to ensure that we have enough redshift bins after the Eulerian to Lagrangian mapping summarized in Figs 2 and 3. We have verified in an additional run A_Δ using 10 redshift bins for the same redshift range that we get the same results as considering bias interpolation. In such a case we have ensured that the redshift bins were wide enough to have only three populations of tracers, i.e. galaxies, which have not changed redshift bin after doing the Eulerian redshift-space to Lagrangian real-space mapping, and galaxies coming from higher and lower redshift bins (see Fig. 6). This can be seen in Fig. 7, where the corresponding populations of galaxies are depicted in different colours after 70 Gibbs-sampling iterations. We can find that the angular response function also has to take into account the different populations according to the displacement field as shown in Fig. 8. The corresponding large-scale tracers are overplotted as red dots. It is interesting to make the visual inspection and verify that the galaxies are on top of the non-vanishing completeness regions, which predict where those galaxies are actually expected to be mapped to according to the same set of displacement fields on the light-cone for a given iteration (see mapping procedure described in Section 3.3 and represented in Figs 2 and 3). The combination of different tracers is done assuming that they are independent tracers of the large-scale structure (without mixed terms), according to the Poisson likelihood as described in the appendix of Ata et al. (2015). This framework permits to add as many tracers of the large-scale structure as additive terms in the log-likelihood used in the posterior PDF within the Hamiltonian sampler. Each of these tracers will have its own bias and response function. We will show how to add different galaxy surveys using this formalism in a forthcoming paper applied to the Cosmological Evolution Survey (COSMOS) field (Laigle et al. 2016; Ata et al., in preparation) and to the local Universe (Kitaura et al., in preparation).

Since the run A with bias interpolation to the position of each galaxy yields numerical identical results to the classification in run

[5] https://www.cosmosim.org/cms/simulations/bigmdpl/

[6] For visualization purposes we also show results with resolutions of $6.25\ h^{-1}$ Mpc, which will be analysed in detail in forthcoming publications.

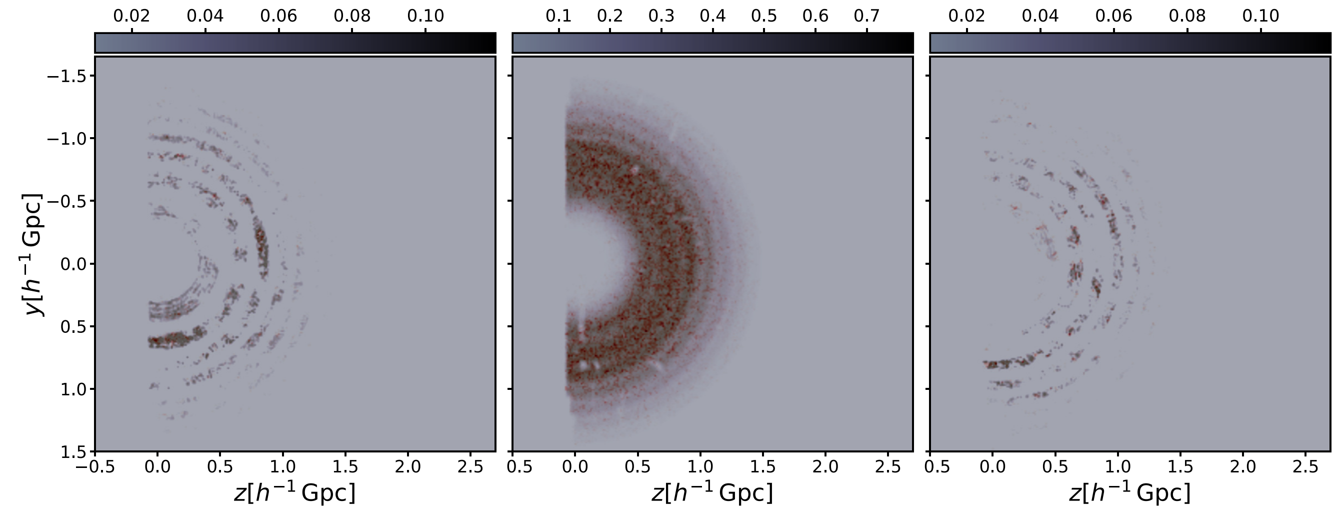

Figure 8. For slices of thickness ∼60 h^{-1} Mpc in the z–y plane of the three-dimensional cubical mesh of side 3200 h^{-1} Mpc and 256^3 cells: completeness for the galaxies that jump to a lower redshift bin (left), for those that stay at the same redshift bin (middle), and for those that jump to an upper redshift bin. Galaxies are overplotted as red dots.

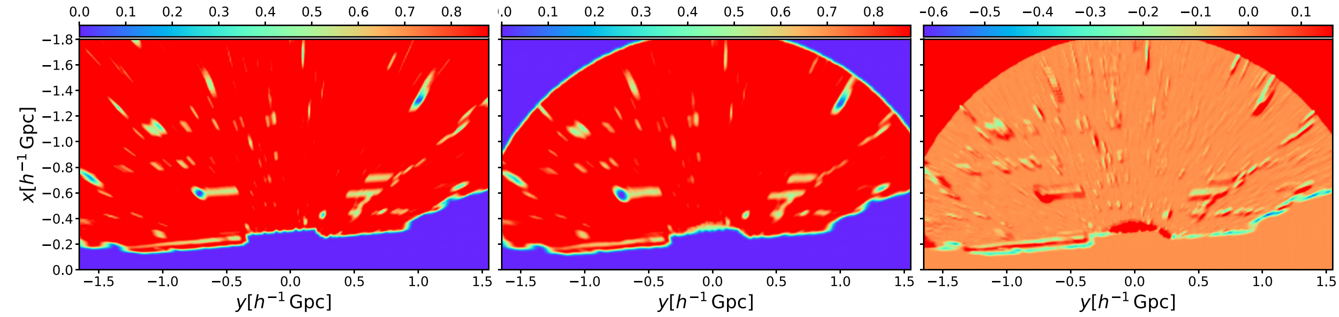

Figure 9. For the same cut as in Fig. 8, but in the x–y plane: 3D projected angular survey geometry, including veto mask in Eulerian space (left), Lagrangian space (middle), and the difference between both (right).

A_Δ, we will from now on consider only variations on run A. We have in this case a single angular mask for all objects, which is mapped from Eulerian redshift space to Lagrangian real space as shown in Fig. 9. Here we can see that edges of the survey become less sharp due to the displacement field. In fact a hole in the survey mask may be displaced or even filled with large-scale structure tracers when going to Lagrangian space. The reconstructed DMDF on the light-cone is shown in Figs 10–12. A first visual inspection shows a substantial correlation between the mock galaxy distribution and the underlying DMDF. The panels in Fig. 10 show the average over 2000 Gibbs-sampling iterations, which are equivalent to about 20 independent samples according to our study shown below. We can see how the structures cancel out in regions far away from the data (red dots). The lower panels in that figure clearly show a 'donut' structure that the reconstruction was performed in a limited redshift range to save computations. The left-hand panels in Fig. 11 show the rich cosmic web for one reconstruction after 70 Gibbs-sampling iterations on a higher resolution of 6.25 h^{-1} Mpc. The right-hand panels show the corresponding Gaussian field, which does not show a transition from the observed to the unobserved region. To get an overview of the calculations done in the COSMIC BIRTH code and to further assess its performance, we show in Fig. 12: the DMDF from the original simulation on the light-cone, but without applying radial selection criteria (upper left-hand panel), the corresponding total response function (upper right-hand panel), the corresponding mock galaxy catalogue in Eulerian redshift space (second row left-hand panel) and in Lagrangian real space (second row right-hand panel), the corresponding reconstructed DMDFs on the light-cone with low (left) and high (right) resolution (panels in the third row), and finally the reconstructed dark matter field at a single snapshot at redshift 100: mean over 2000 iterations and one reconstructed sample after 1000 iterations. The panels in the second row show how the galaxy distribution becomes considerably more homogeneous after reconstruction. The lower panels in Fig. 12 show that neither survey nor radial selection effects are present in the reconstruction, the data region is not distinguished in the single reconstruction (right-hand panel), while the ensemble average clearly shows an enlarged region (right-hand panel in the second row) of the original data region (left-hand panel in the second row) due to the action of gravity.

To make a quantitative assessment we compute the mean and variance of the power spectra of the reconstructed density fields at $z = 100$, as shown in the right-hand panels of Fig. 13. The upper panel shows that we obtain exquisite unbiased power spectra including tidal fields and a non-linear small-scale correction (ALPT) with a full light-cone treatment. Also the statistics of the reconstructed density fields are Gaussian. We find that the particular realization of the BigMDPL simulation has a slight excess of power on large scales

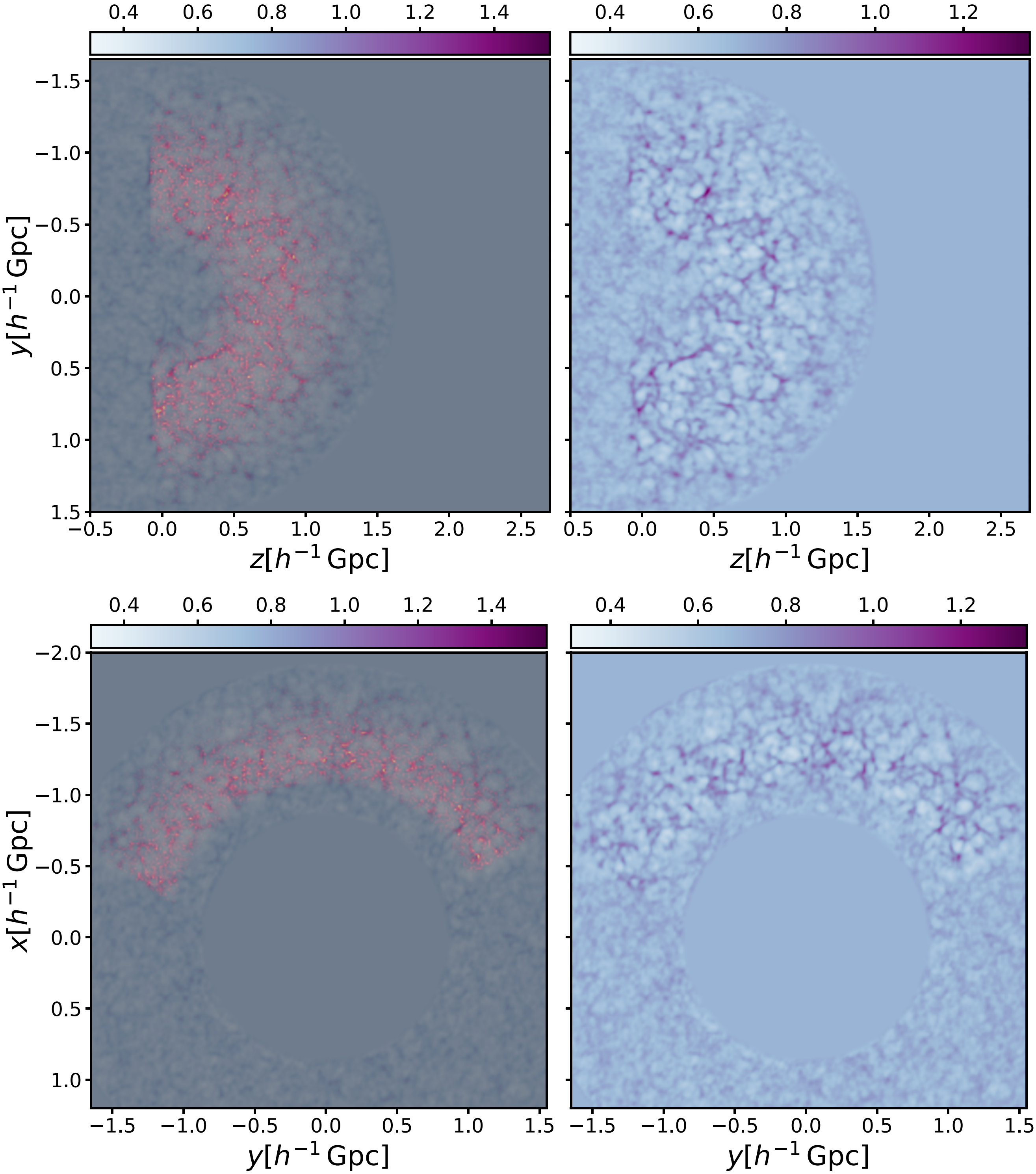

Figure 10. Slices of thickness ∼60 h^{-1} Mpc in the *z*–*y* (top) and *x*–*y* (bottom) plane of the three-dimensional cubical mesh of side 3200 h^{-1} Mpc and 256^3 cells showing the averaged density field over 20 independent samples. Left-hand panels with (right-hand panels without) galaxies overplotted.

that is accurately reproduced in the reconstructions (see Klypin et al. 2016). The lower right-hand panel shows that the convergence of the COSMIC BIRTH code is extremely fast. It requires only about 30 iterations to converge within percentage accuracy to the theoretical power spectrum and quickly gets the right shape after about 10 iterations (see also Appendix B).

We have performed two additional runs B and C. In run B shown in the lower left-hand panel of Fig. 13, we used a normal cascaded integrator–comb (CIC) interpolation scheme (without tetrahedral tessellation and Lagrangian tetrahedral tessellation) to construct the angular response function in Lagrangian space with the subset of the 256^3 cells enclosed in the considered redshift range. This

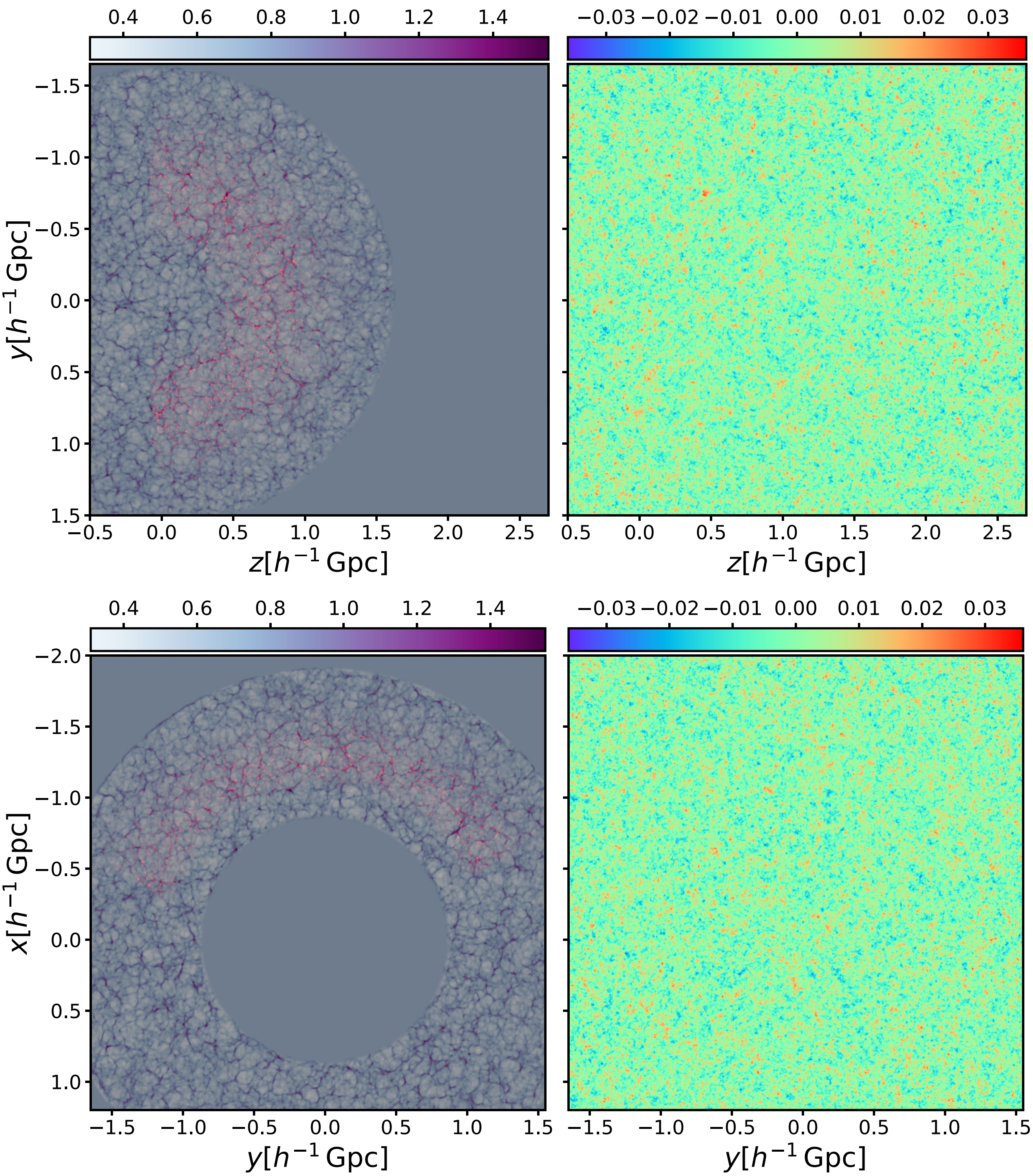

Figure 11. Left-hand panels: same as Fig. 10, but for single reconstructions after 70 Gibbs-sampling iterations on meshes with 512^3 cells. Right-hand panels: corresponding initial density fields at $z = 100$ are shown.

inaccuracy given the low number of tracers used to do the Eulerian–Lagrangian mapping had the same effect as a radial selection function as can be seen in the abnormal excess of power on large scales. In fact, using a single redshift bin for the whole CMASS data had a similar effect as found in run C shown in the upper left-hand panel.

Another quantitative measurement of the speed of the COSMIC BIRTH code is presented in the upper left-hand panel of Fig. 14. Here we demonstrate that the correlation length is of about 100 iterations, meaning that we get independent samples with considerably lower number of iterations between samples than previous methods (see Section 1).

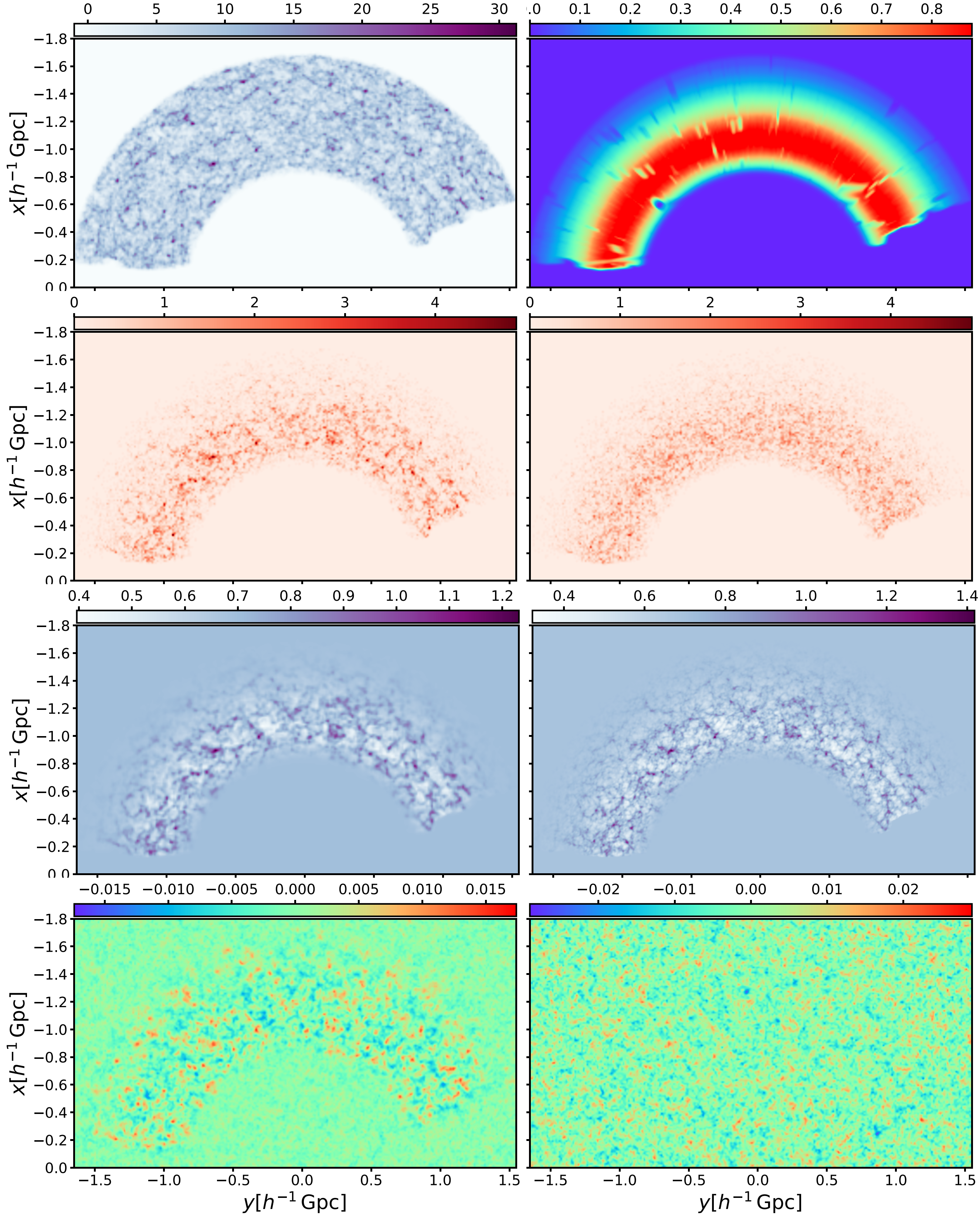

Figure 12. For the same cut as the lower panels in Fig. 10. Upper left: dark matter from the BigMD simulation. Upper right: CMASS completeness. Second row: galaxy number counts based on the BigMD simulation using SHAM in Eulerian (left) and Lagrangian (right) space (reconstructed). Third row: dark matter reconstructions with ALPT on 256^3 (left) and 512^3 (right). Fourth row: reconstructed density field at $z = 100$, average (left), single (right).

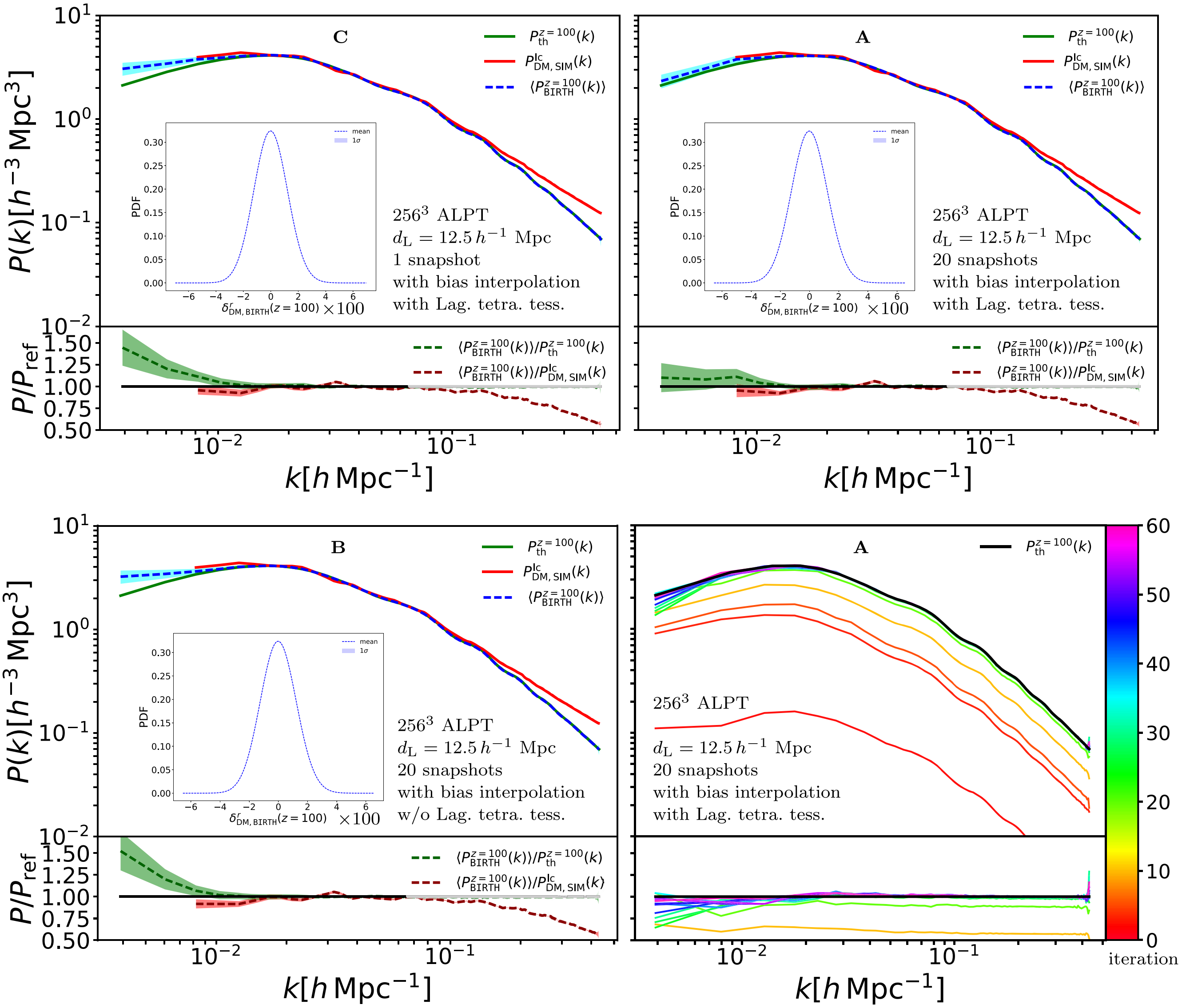

Figure 13. Statistics of the reconstructed density fields at $z = 100$ from Bayesian posterior sampling with a lognormal-Poisson model for Lagrangian tracers. The upper right-hand panel shows results with 20 redshift snapshots (run A), the left one with only one (run C), both using ALPT. The lower left-hand panel corresponds to a run with 20 redshift snapshots using ALPT without Lagrangian tetrahedral tessellation (Lag. tetra. tess.; run B), while the lower right-hand panel shows the convergence of the run A (the colour bar indicates the iteration number). The mean is represented by the blue dashed curve with the corresponding 1σ region in cyan, both for the power spectrum and the matter PDF (skewness and kurtosis are of the order of 10^{-2} and 10^{-4}, respectively). The theoretical mean power spectrum is represented by the solid green line. The measured power spectrum from the light-cone DMDF normalized to a mean redshift of $z = 0.57$ is in red. The corresponding ratio power spectra are presented at the bottom of each panel. The red solid line is close to the green one because the BigMD simulation was run with initial conditions with a corresponding power spectrum relatively close to the theoretical one, selected from a set of random Gaussian realizations.

The correlation length of the density bins over the iteration distance is calculated as

$$C_n(\sigma_j) = \frac{1}{N-n}\sum_{i=0}^{N-n}\frac{\left(\delta_j^i - \langle\delta_j\rangle\right)\left(\delta_j^{i+n} - \langle\delta_j\rangle\right)}{\sigma^2(\delta_j)}, \tag{21}$$

where δ_j is the overdensity field in each iteration at voxel j, N is the number of samples, and n is the distance between iterations.

We have computed a series of cross-correlations following the definitions in Kitaura et al. (2012b) and Heß et al. (2013) shown in Fig. 14. The analysis is restricted to $k = 0.2\,h\,{\rm Mpc}^{-1}$, which is about 50 per cent of the Nyquist frequency given the considered mesh resolution. The lower left-hand panel shows that after about 2000 iterations the mean over reconstructed dark matter fields on the redshift space light-cone cross-correlated with the corresponding mock galaxy distribution does not improve, meaning that the ensemble average can be considered to have converged after only 2000 iterations. The right-hand panel shows that we get close to the optimal cross-correlation achieved between the real-space dark matter field and the galaxy field on a full box without selection criteria (cyan line), as with our reconstruction (dashed dotted black line), which is a considerable information gain with respect to the red line when including the same selection criteria from the galaxy field in the dark matter field. This means that the Bayesian code is actually correcting

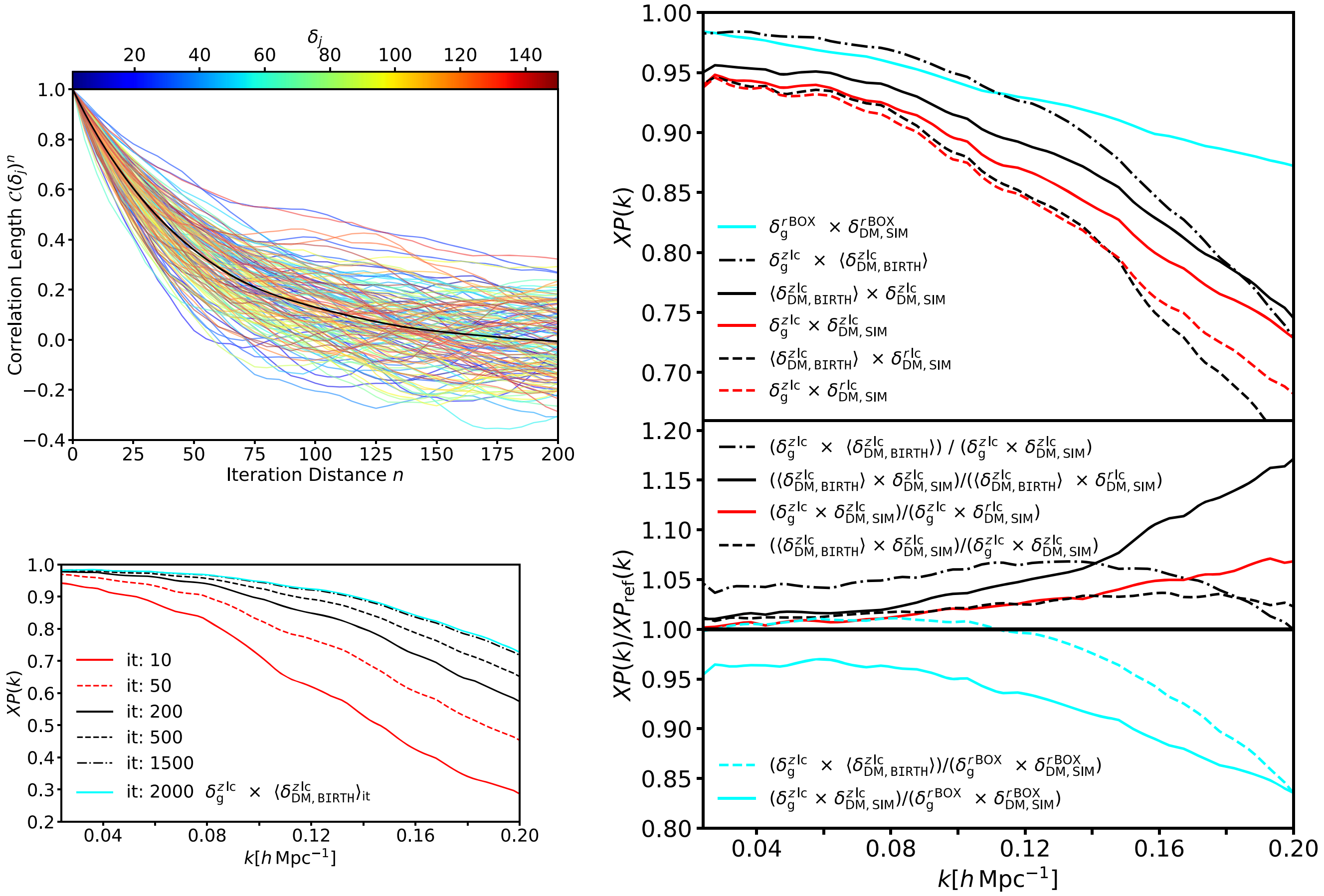

Figure 14. Convergence behaviour and assessment of the information gain from the Bayesian posterior. Upper left-hand panel: correlation length with the mean over all density bins for each iteration represented by a solid black line. This demonstrates that independent samples are drawn each ∼100 iterations considering the density field. Lower left-hand panel: growth of the power spectrum with increasing iteration number. After ∼40 iterations the reconstructed power spectrum gains its target amplitude, fluctuating according to the statistical uncertainty of the individual realizations. The right-hand panel presents cross-power spectra (top) and the corresponding ratios (bottom) between different combinations of fields, as shown in the legend. The superscript 'lc' stands for light-cone, if it is accompanied by r or by z in real space or in redshift space, respectively. The superscript BOX stands for data occupying the entire box without survey geometry, nor radial selection effects.

with the given structure formation model and the response function for incompleteness. This information gain is of course lost towards small scales and the dashed–dotted line drops towards high k. It is in fact remarkable, how well the galaxy and reconstructed dark matter field correlate with each other given that a whole structure formation model displaces the large-scale structure tracers on average 8–10 h^{-1} Mpc. This is trivial when the dark matter field is simply a smooth version of the galaxy field, which is far from being the case here. It is also interesting to verify that the cross-correlation between the dark matter reconstructions and the true dark matter field is larger than the cross-correlation between the mock galaxy field and the dark matter field. This means that our actual structure formation and galaxy bias modelling is working. We also find a considerably higher correlation between the dark matter field reconstruction in redshift space and the true simulated dark matter field in redshift space than in real space (solid black line versus dashed black line). This implies that the redshift-space distortions modelling is meaningful. The lower right-hand panel shows the information gain. Here we find consistent results demonstrating that the reconstructions are adding information through the structure formation, bias, and completeness modelling.

5 DISCUSSION AND CONCLUSIONS

In this work, we have presented the COSMIC BIRTH method. It provides a Bayesian framework to tackle the matter reconstruction problem from a distribution of galaxies.

It is a particularly simple and efficient algorithm, which solves the Bayesian reconstruction problem including selection effects and non-linear structure formation in the calculation of the displacements. It is important to stress that this is achieved without giving-up on accuracy or loss of generality. The strategy of splitting the approach into two reconstructions steps permits us to use a lognormal-Poisson posterior in Lagrangian space, as this model is accurate at initial cosmic times (high redshifts). The lognormal assumption ensures that the density field is positive definite converging to the Gaussian assumption for $|\delta| \ll 1$. The Poisson distribution function permits us to correct for aliasing caused by describing the galaxy distribution (in Lagrangian space) as discrete number counts of large-scale structure tracers on a regular mesh. Note that a Gaussian likelihood keeping only the two-point statistics of the Poisson likelihood is not adaptive, but yields a constant mean noise covariance matrix (see Kitaura et al. 2009, 2010). This limits very much the accuracy of Wiener filtering based on a Gaussian prior and a Gaussian likelihood (Zaroubi et al.

1995). Therefore we conclude that the simplest statistical model we can consider is the lognormal-Poisson one. Thanks to this model (positive definite matter fields connected to a discrete number of tracers), we can include a non-linear bias description beyond the commonly used linear one in BAO reconstruction. This Lagrangian posterior model yields the primordial Gaussian density fields defined on a regular mesh assuming a set of observed Lagrangian tracers using HMC sampling. These Lagrangian tracers in turn are connected to the observed galaxy sample on the light-cone through a forward modelling within an iterative Gibbs-sampling scheme based on an arbitrary structure formation model. This approach dramatically simplifies the programming structure of the code, as no gradients of structure formation models need to be computed. In this way, the structure formation model can be changed by any other one in this framework, and only needs to deliver information on the initial and final positions of tracers including their peculiar velocities. We will investigate in a subsequent work how a particle mesh code improves the results (this was investigated to some extent with the KIGEN code going from the second-order Lagrangian perturbation theory (2LPT) to ALPT; Kitaura et al. 2012b; Heß et al. 2013). The COSMIC BIRTH method combines a grid-based with a tracer-based reconstruction, relying on Bayesian inference methods while directly yielding a set of Lagrangian tracers equivalent to BAO reconstruction.

We have introduced technical improvements to the Hamiltonian-sampling scheme to gain a factor of about 20 in efficiency, yielding correlation lengths of only 40–50 iterations. This demonstrates that Bayesian methods can actually be practically used to sample posterior PDFs. Part of these improvements are further studied in detail in a companion paper and are inspired by techniques widely used in lattice quantum field theory using higher order discretizations of the Hamiltonian equations of motions (Hernández-Sánchez et al. 2019). Furthermore, we have introduced a strategy to efficiently deal with non-diagonal Hamiltonian mass matrices including complex survey geometries, which speeds up the convergence by up to ∼70 per cent with high acceptance rates of 60–70 per cent. The idea is based on associating specific uncertainties in the data augmentation modelled by the momenta depending on the completeness. Therefore, the Hamiltonian mass needs to include the structure of the response function as derived in Appendix B. Special attention must be paid to a consistent formulation of the square root of the Hamiltonian mass, as required to generate the random fluctuations, and its inverse to solve Hamiltonian's equations of motion.

This approach has the novelty of being only dependent on cosmological parameters and an arbitrary structure formation model, while solving the problem of dealing with galaxy bias on the light-cone. As we have shown, one needs in general (and in particular for the BOSS data) to consider a varying galaxy bias with redshift (see Fig. 5 and Kitaura et al. 2016a). However, one can certainly find unbiased power spectra with respect to the theoretical one in the full cubical volume with a single (wrong) bias parameter, which as we know now is inaccurate (see e.g. Ata et al. 2017). These kinds of crude approximations will have an impact in a detailed tomographic analysis, and will not permit to break degeneracies with e.g. gravity- or neutrino-induced deviations in the power spectrum. We made progress to include a complete robust non-linear Lagrangian bias framework, and a Lagrangian tetrahedral tessellation of the survey geometry from Eulerian to Lagrangian coordinates, as it is required in our framework. Further investigation in this direction will permit us to better understand Lagrangian bias. Also this method can be used to study Eulerian bias from the data itself and the dark matter reconstructions, without having used any Eulerian bias description in the reconstruction process. In this work, we have found a connection between the measurable large-scale bias and the effective non-linear bias in Lagrangian space, solving the dependence on the mesh resolution. In this way non-local bias is accounted for through the displacement field.

The method has the potential to become a standard technique (particularly for BAO reconstruction, as we will show in a subsequent paper). Our tests demonstrate that we can obtain unbiased dark matter field reconstructions on the light-cone from highly biased tracers using arbitrary structure formation models. Therefore, this method shows its great potential for the analysis of deep redshift surveys such as DESI, *Euclid*, J-PAS, PFS, *WFIRST*, 4MOST, etc. Provided its sampling speed, other more general applications can be foreseen with this method, such as cosmological parameter estimation and growth rate sampling. We expect that this method contributes towards a full analysis of the large-scale structure, ultimately including a full determination of the cosmological model.

ACKNOWLEDGEMENTS

We thank the referee for the careful revision of the manuscript. The authors thank Raúl E. Angulo, Ginevra Favole, Mariana Vargas-Magaña, and Cheng Zhao for discussions. F-SK acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (MINECO) under the Severo Ochoa program SEV-2015-0548, and for the grants RYC2015-18693 and AYA2017-89891-P. SAR-T thanks support from grants SEV-2015-0548 and RAYA2017-89891-P. MA thanks for the hospitality at the IAC and support from the Kavli IPMU fellowship. MH-S acknowledges the *residente* grant at IAC. AB-A acknowledges financial support from the Spanish Ministry of Economy and Competitiveness (MINECO) under the Severo Ochoa program SEV-2015-0548. GY acknowledges financial support from MINECO/FEDER under research grant AYA2015-63810-P and MICIU/FEDER PGC2018-094975-C21. F-SK thanks Joaquín Pérez Macía for studying the efficiency of the non-diagonal Hamiltonian mass matrix during his bachelor thesis.

DATA AVAILABILITY

The data underlying this paper will be shared on reasonable request to the corresponding author.

REFERENCES

Abel T., Hahn O., Kaehler R., 2012, MNRAS, 427, 61
Abidi M. M., Baldauf T., 2018, J. Cosmol. Astropart. Phys., 07, 029
Ahn K., Iliev I. T., Shapiro P. R., Srisawat C., 2015, MNRAS, 450, 1486
Akeson R. et al., 2019, preprint (arXiv:1902.05569)
Alam S. et al., 2017, MNRAS, 470, 2617
Amendola L. et al., 2018, Living Rev. Relativ., 21, 2
Anderson L. et al., 2014, MNRAS, 441, 24
Ata M., Kitaura F.-S., Müller V., 2015, MNRAS, 446, 4250
Ata M. et al., 2017, MNRAS, 467, 3993
Ata M. et al., 2018, MNRAS, 473, 4773
Aubourg É. et al., 2015, Phys. Rev. D, 92, 123516
Balaguera-Antolínez A., Bilicki M., Branchini E., Postiglione A., 2018, MNRAS, 476, 1050
Balaguera-Antolínez A., Kitaura F.-S., Pellejero-Ibáñez M., Zhao C., Abel T., 2019, MNRAS, 483, L58
Balaguera-Antolínez A. et al., 2020, MNRAS, 491, 2565
Behroozi P. S., Wechsler R. H., Wu H.-Y., 2013, ApJ, 762, 109
Benitez N. et al., 2014, preprint (arXiv:1403.5237)
Birkin J., Li B., Cautun M., Shi Y., 2019, MNRAS, 483, 5267
Blake C., Glazebrook K., 2003, ApJ, 594, 665

Bolton A. S. et al., 2012, AJ, 144, 144
Bond J. R., Kofman L., Pogosyan D., 1996, Nature, 380, 603
Bos E. G. P., Kitaura F.-S., van de Weygaert R., 2019, MNRAS, 488, 2573
Brenier Y., Frisch U., Hénon M., Loeper G., Matarrese S., Mohayaee R., Sobolevskiuzi A., 2003, MNRAS, 346, 501
Cen R., Ostriker J. P., 1992, ApJ, 399, L113
Coles P., Jones B., 1991, MNRAS, 248, 1
Conway E. et al., 2005, MNRAS, 356, 456
Courtois H. M., Pomarède D., Tully R. B., Hoffman Y., Courtois D., 2013, AJ, 146, 69
Cresswell J. G., Percival W. J., 2009, MNRAS, 392, 682
Creutz M., Gocksch A., 1989, Phys. Rev. Lett., 63, 9
Crocce M., Scoccimarro R., 2008, Phys. Rev. D, 77, 023533
Dawson K. S. et al., 2013, AJ, 145, 10
Dawson K. S. et al., 2016, AJ, 151, 44
de Jong R. S. et al., 2019, The Messenger, 175, 3
de la Torre S., Peacock J. A., 2013, MNRAS, 435, 743
Desjacques V., Jeong D., Schmidt F., 2018, Phys. Rep., 733, 1
Duane S., Kennedy A., Pendleton B. J., Roweth D., 1987, Phys. Lett. B, 195, 216
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007, ApJ, 664, 675
Falck B. L., Neyrinck M. C., Szalay A. S., 2012, ApJ, 754, 126
Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, MNRAS, 463, 2273
Fry J. N., 1996, ApJ, 461, L65
Fry J. N., Gaztanaga E., 1993, ApJ, 413, 447
Gil-Marín H., Noreña J., Verde L., Percival W. J., Wagner C., Manera M., Schneider D. P., 2015, MNRAS, 451, 539
Gil-Marín H., Percival W. J., Verde L., Brownstein J. R., Chuang C.-H., Kitaura F.-S., Rodríguez-Torres S. A., Olmstead M. D., 2017, MNRAS, 465, 1757
Granett B. R. et al., 2015, A&A, 583, A61
Gunn J. E. et al., 2006, AJ, 131, 2332
Guzzo L. et al., 2008, Nature, 451, 541
Guzzo L. et al., 2014, A&A, 566, A108
Hada R., Eisenstein D. J., 2019, MNRAS, 482, 5685
Hahn O., Abel T., Kaehler R., 2013, MNRAS, 434, 1171
Hamilton A. J. S., Tegmark M., 2004, MNRAS, 349, 115
Heath D. J., 1977, MNRAS, 179, 351
Heß S., Kitaura F.-S., Gottlöber S., 2013, MNRAS, 435, 2065
Hernández-Sánchez M., Kitaura F.-S., Ata M., Vecchia C. D., 2019, preprint (arXiv:1911.02667)
Horowitz B., Seljak U., Aslanyan G., 2019, J. Cosmol. Astropart. Phys., 10, 035
Ishak M., 2019, Living Rev. Relativ., 22, 1
Jasche J., Kitaura F. S., 2010, MNRAS, 407, 29
Jasche J., Lavaux G., 2019, A&A, 625, A64
Jasche J., Wandelt B. D., 2013, MNRAS, 432, 894
Kaiser N., 1984, ApJ, 284, L9
Kaiser N., 1987, MNRAS, 227, 1
Kitaura F.-S., 2013, MNRAS, 429, L84
Kitaura F.-S., Angulo R. E., 2012, MNRAS, 425, 2443
Kitaura F. S., Enßlin T. A., 2008, MNRAS, 389, 497
Kitaura F.-S., Hess S., 2013, MNRAS, 435, L78
Kitaura F. S., Jasche J., Li C., Enßlin T. A., Metcalf R. B., Wand elt B. D., Lemson G., White S. D. M., 2009, MNRAS, 400, 183
Kitaura F.-S., Jasche J., Metcalf R. B., 2010, MNRAS, 403, 589
Kitaura F.-S., Gallerani S., Ferrara A., 2012a, MNRAS, 420, 61
Kitaura F.-S., Erdoğdu P., Nuza S. E., Khalatyan A., Angulo R. E., Hoffman Y., Gottlöber S., 2012b, MNRAS, 427, L35
Kitaura F.-S., Yepes G., Prada F., 2014, MNRAS, 439, L21
Kitaura F.-S., Gil-Marín H., Scóccola C. G., Chuang C.-H., Müller V., Yepes G., Prada F., 2015, MNRAS, 450, 1836
Kitaura F.-S. et al., 2016a, MNRAS, 456, 4156
Kitaura F.-S., Ata M., Angulo R. E., Chuang C.-H., Rodríguez-Torres S., Monteagudo C. H., Prada F., Yepes G., 2016b, MNRAS, 457, L113
Klypin A., Yepes G., Gottlöber S., Prada F., Heß S., 2016, MNRAS, 457, 4340
Laigle C. et al., 2016, ApJS, 224, 24
Leclercq F., Jasche J., Wandelt B., 2015, J. Cosmol. Astropart. Phys., 06, 015
Levi M. et al., 2013, preprint (arXiv:1308.0847)
Libeskind N. I. et al., 2018, MNRAS, 473, 1195
Lindsay S. N. et al., 2014, MNRAS, 440, 1527
LSST Science Collaboration, 2009, preprint (arXiv:0912.0201)
Ludlow A. D., Porciani C., 2011, MNRAS, 413, 1961
McDonald P., Roy A., 2009, J. Cosmol. Astropart. Phys., 08, 020
Martel H., 2005, preprint (arXiv:astro-ph/0506540)
Mirbabayi M., Schmidt F., Zaldarriaga M., 2015, J. Cosmol. Astropart. Phys., 07, 030
Modi C., Castorina E., Seljak U., 2017, MNRAS, 472, 3959
Monaco P., Efstathiou G., 1999, MNRAS, 308, 763
Neal R. M., 2012, preprint (arXiv:1206.1901)
Neyrinck M. C., 2012, MNRAS, 427, 494
Neyrinck M. C., 2013, MNRAS, 428, 141
Neyrinck M. C., Aragón-Calvo M. A., Jeong D., Wang X., 2014, MNRAS, 441, 646
Nusser A., Branchini E., 2000, MNRAS, 313, 587
Nusser A., Davis M., 1994, ApJ, 421, L1
Nusser A., Davis M., Branchini E., 2014, ApJ, 788, 157
Nuza S. E., Kitaura F.-S., Heß S., Libeskind N. I., Müller V., 2014, MNRAS, 445, 988
Padmanabhan N., Xu X., Eisenstein D. J., Scalzo R., Cuesta A. J., Mehta K. T., Kazin E., 2012, MNRAS, 427, 2132
Pan H., Obreschkow D., Howlett C., Lagos C. d. P., Elahi P. J., Baugh C., Gonzalez-Perez V., 2020, MNRAS, 493, 747
Peebles P. J. E., 1980, The Large-Scale Structure of the Universe. Princeton Univ. Press, Princeton, NJ
Peebles P. J. E., 1989, ApJ, 344, L53
Pellejero-Ibañez M. et al., 2020, MNRAS, 493, 586
Percival W. J., Schäfer B. M., 2008, MNRAS, 385, L78
Percival W. J., Cole S., Eisenstein D. J., Nichol R. C., Peacock J. A., Pope A. C., Szalay A. S., 2007, MNRAS, 381, 1053
Perlmutter S. et al., 1999, ApJ, 517, 565
Platen E., van de Weygaert R., Jones B. J. T., Vegter G., Calvo M. A. A., 2011, MNRAS, 416, 2494
Reid B. et al., 2016, MNRAS, 455, 1553
Riess A. G. et al., 1998, AJ, 116, 1009
Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, 1173
Sarpa E., Schimd C., Branchini E., Matarrese S., 2019, MNRAS, 484, 3818
Saslaw W. C., 1989, ApJ, 341, 588
Schmidt F., Jeong D., Desjacques V., 2013, Phys. Rev. D, 88, 023515
Schmidt F., Elsner F., Jasche J., Nguyen N. M., Lavaux G., 2019, J. Cosmol. Astropart. Phys., 01, 042
Schmittfull M., Feng Y., Beutler F., Sherwin B., Chu M. Y., 2015, Phys. Rev. D, 92, 123522
Seljak U., Yu B., 2019, preprint (arXiv:1901.04454)
Seljak U. et al., 2005, Phys. Rev. D, 71, 043511
Seo H.-J., Eisenstein D. J., 2003, ApJ, 598, 720
Shandarin S. F., Zel'dovich Y. B., 1989, Rev. Modern Phys., 61, 185
Shandarin S., Habib S., Heitmann K., 2012, Phys. Rev. D, 85, 083005
Sheth R. K., 1998, MNRAS, 299, 207
Sheth R. K., Chan K. C., Scoccimarro R., 2013, Phys. Rev. D, 87, 083002
Shi Y., Cautun M., Li B., 2018, Phys. Rev. D, 97, 023505
Slepian Z. et al., 2017, MNRAS, 469, 1738
Smee S. A. et al., 2013, AJ, 146, 32
Sorce J. G., Courtois H. M., Gottlöber S., Hoffman Y., Tully R. B., 2014, MNRAS, 437, 3586
Springel V., 2005, MNRAS, 364, 1105
Swanson M. E. C., Tegmark M., Hamilton A. J. S., Hill J. C., 2008, MNRAS, 387, 1391
Takada M. et al., 2014, PASJ, 66, R1
Tanner M. A., 1993, Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 2nd edn. Springer-Verlag, New York

Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, J. Cosmol. Astropart. Phys., 06, 036

Taylor J. F., Ashdown M. A. J., Hobson M. P., 2008, MNRAS, 389, 1284

Tegmark M., Peebles P. J. E., 1998, ApJ, 500, L79

The Dark Energy Survey Collaboration, 2005, preprint (arXiv:astro-ph/0510346)

Tully R. B., Courtois H., Hoffman Y., Pomarède D., 2014, Nature, 513, 71

Verde L. et al., 2002, MNRAS, 335, 432

Wang H., Mo H. J., Yang X., van den Bosch F. C., 2012, MNRAS, 420, 1809

Wang H., Mo H. J., Yang X., van den Bosch F. C., 2013, ApJ, 772, 63

Wang H., Mo H. J., Yang X., Jing Y. P., Lin W. P., 2014, ApJ, 794, 94

White M., 2014, MNRAS, 439, 3630

Yahil A., Strauss M. A., Davis M., Huchra J. P., 1991, ApJ, 372, 380

Zaroubi S., Hoffman Y., Fisher K. B., Lahav O., 1995, ApJ, 449, 446

Zaroubi S., Hoffman Y., Dekel A., 1999, ApJ, 520, 413

Zel'dovich Y. B., 1970, A&A, 5, 84

Zhao G.-B. et al., 2017, Nat. Astron., 1, 627

APPENDIX A: RENORMALIZED PERTURBATION THEORY

Here we derive the third-order equation from renormalized perturbation theory, which connects the non-linear bias correction with the large-scale bias and the dark matter variance at the cell resolution. Let us perform a Taylor expansion in the expression to third order in the overdensity field:

$$\begin{aligned}\delta_{\rm g}(z) \equiv \frac{\rho_{\rm g}(z)}{\bar{N}} - 1 \simeq \tau(z)\Bigg[1 + b(z) f_{\rm b}(z)\delta(z) \\ + \frac{1}{2} b(z) f_{\rm b}(z)(b(z) f_{\rm b}(z) - 1)\,(\delta(z))^2 \\ + \frac{1}{3!} b(z) f_{\rm b}(z)(b(z) f_{\rm b}(z) - 1)(b(z) f_{\rm b}(z) - 2)\,(\delta(z))^3\Bigg] - 1, \quad \text{(A1)}\end{aligned}$$

with $\tau(z) \equiv \gamma(z)/\bar{N}$. The usual expression for the perturbatively expanded overdensity field to third order ignoring non-local terms is given by (see e.g. Desjacques et al. 2018)

$$\delta_{\rm g}(z) = c_\delta(z)\delta(z) + \frac{1}{2} c_{\delta^2}(z)(\delta^2(z) - \sigma^2(z)) + \frac{1}{3!} c_{\delta^3}(z)\delta^3(z). \quad \text{(A2)}$$

Correspondingly, one can show that the large-scale bias is given by (e.g. McDonald & Roy 2009)

$$b_\delta(z) = c_\delta(z) + \frac{34}{21} c_{\delta^2}(z)\sigma^2(z) + \frac{1}{2} c_{\delta^3}(z)\sigma^2(z). \quad \text{(A3)}$$

By considering that in our case the large-scale bias is given by $b(z)$ and identifying the coefficients $\{c_\delta = \tau\, f_{\rm b} b, c_{\delta^2} = \tau\, f_{\rm b} b(f_{\rm b} b - 1), c_{\delta^3} = \tau\, f_{\rm b} b(f_{\rm b} b - 1)(f_{\rm b} b - 2), \tau - 1 = -c_{\delta^2}\sigma^2/2\}$ from equations (A1) and (A2) one can derive the following cubic equation for $f_{\rm b}$:

$$\begin{aligned}b(z)\, f_{\rm b}^3(z) + f_{\rm b}(z)^2\left(\frac{5}{21} - b(z)\right) \\ + \frac{f_{\rm b}(z)}{b(z)}\left(\frac{2}{\sigma^2(z)} - \frac{26}{21} + b(z)\right) - \frac{2}{\sigma^2(z)\, b(z)} = 0. \quad \text{(A4)}\end{aligned}$$

We have verified that this model yields accurate power spectra on large scales, as long as the bias is given by the truncated Taylor expansion at third order. Although the absolute value of the overdensity field is smaller than one at high redshift (say $z = 100$) and resolutions of a few Mpc (say about 5 h^{-1} Mpc), the Lagrangian large-scale bias is so high that higher order terms in the Taylor expansion are still relevant. Thus the validity range of this framework is thus restricted to special cases, such as low bias tracers.

APPENDIX B: EFFICIENT NON-DIAGONAL HAMILTONIAN MASS

The HMC method (Duane et al. 1987) requires a nuisance variable to sample the PDF, which is called the momenta $\boldsymbol{p}$. According to the mechanical analogy the kinetic energy is given by

$$K(\boldsymbol{p}|\mathbf{M}) = \frac{1}{2}\boldsymbol{p}^{\rm t}\mathbf{M}^{-1}\boldsymbol{p}, \quad \text{(B1)}$$

where $\mathbf{M}$ is the Hamiltonian mass, which acts as a pre-conditioner of the HMC sampler, and can considerably speed up the HMC (Neal 2012). This mass can be interpreted as the covariance matrix of the momenta. The kinetic term can be connected to a multivariate Gaussian distribution proportional to $\exp[-K]$. This implies that the generation of the momenta is equivalent to the generation of a Gaussian field with an appropriate covariance matrix $\mathbf{M}$. Ideally, this mass should have the structure of the prior and of the likelihood (Jasche & Kitaura 2010), i.e. a term related to the matter field covariance matrix, say $\mathbf{C}$, and a term related to the response function $\mathbf{R}$, which can have a structure like this:

$$\mathbf{M} = \mathbf{C}^{-1} + \beta\mathbf{R}, \quad \text{(B2)}$$

where β is a constant that will depend on the number density and maybe other quantities. This mass represents a non-diagonal matrix in neither Fourier nor configuration space, as $\mathbf{C}$ is diagonal in Fourier space, but $\mathbf{R}$ is diagonal in real space, being the three-dimensional completeness. For an analysis without selection function, nor angular completeness, i.e. considering a full complete volume, the second term in equation (B2) can be neglected (see Taylor, Ashdown & Hobson 2008). However, it is clear that an efficient sampler needs information on the completeness of the volume in a realistic case, as the uncertainty in our reconstruction is not the same in a well-sampled area, as in unobserved one. Here we face two different problems. One problem is that we need the inverse of the mass matrix, as we need to numerically solve the Hamiltonian equations of motion to perform HMC sampling. In particular, we need to solve this equation involving the mass matrix:

$$\frac{{\rm d}\boldsymbol{x}}{{\rm d}t} = \frac{\partial\mathcal{H}}{\partial\boldsymbol{p}} = \mathbf{M}^{-1}\boldsymbol{p}, \quad \text{(B3)}$$

where $\boldsymbol{x}$ are the positions (in our case the matter density field at initial cosmic times), t cosmic time, and $\mathcal{H}$ the Hamiltonian. One could consider applying efficient inversion schemes based on conjugate gradients (see Kitaura & Enßlin 2008, and references therein), but is clear that this will lower the efficiency of the HMC sampler. But, this is not even so trivial, as we have a second problem, since we also need the square root of the mass matrix $\sqrt{\mathbf{M}}$ to efficiently generate the Gaussian field of momenta in Fourier space (see e.g. Martel 2005). For that reason let us consider a different mass matrix factorizing the term that is diagonal in Fourier space with the one diagonal in configuration space:

$$\mathbf{M} = \mathbf{C}^{-1}(\mathbb{1} + \beta\mathbf{C}\mathbf{R}), \quad \text{(B4)}$$

$$\mathbf{M} \simeq \mathbf{C}^{-1}(\mathbb{1} + c\mathbf{R}), \quad \text{(B5)}$$

approximating $\mathbf{C}$ by a constant inside the parenthesis, which multiplied by β yields an effective constant of c. Computing the inverse of such matrix is trivial now, however, writing its square root is not. Since the mass matrix is a free quantity, let us consider the naive expression for the square root as valid:

$$\sqrt{\mathbf{M}} = \mathbf{C}^{-\frac{1}{2}}(\mathbb{1} + c\mathbf{R})^{\frac{1}{2}}. \quad \text{(B6)}$$

We can now go the other way round and derive the corresponding mass matrix and in particular its inverse:

$$\mathbf{M} = \mathbf{C}^{-\frac{1}{2}} \cdot (\mathbb{1} + c \cdot \mathbf{R})^{\frac{1}{2}} \cdot \mathbf{C}^{-\frac{1}{2}} \cdot (\mathbb{1} + c \cdot \mathbf{R})^{\frac{1}{2}}, \tag{B7}$$

$$\mathbf{M}^{-1} = (\mathbb{1} + c \cdot \mathbf{R})^{-\frac{1}{2}} \cdot \mathbf{C}^{\frac{1}{2}} \cdot (\mathbb{1} + c \cdot \mathbf{R})^{-\frac{1}{2}} \cdot \mathbf{C}^{\frac{1}{2}}. \tag{B8}$$

The important aspect to keep track of is that all expressions of the mass matrix have to be consistent. We have found in this way an expression for the Hamiltonian mass matrix that can be efficiently applied as a series of convolutions going from configuration to Fourier space back and forth both for generating the momenta, where the square root is required, and to solve the Hamiltonian equations of motions, where the inverse is needed. We have performed a series of numerical tests with a fourth-order discretization of the Hamiltonian equations of motions to find the optimal c value between 0.2 and 0.3 given our setting (number density, survey geometry, and radial selection function). As an example a run with $c = 0$ requiring 156 min (70 iterations) of CPU time with eight cores until the power spectra are within 1 per cent compatible with the theoretical one at $k > 0.1\,h\,\mathrm{Mpc}^{-1}$ (to avoid cosmic variance at lower k-values), took 48 min (28 iterations) with $c = 0.2$. Hence we can gain a speed up of about 70 per cent in the convergence of the HMC sampler.

This paper has been typeset from a TEX/LATEX file prepared by the author.